ESTIMATION PROBLEMS IN SPEECH AND NATURAL LANGUAGE

BY

SUMA PALLATHADKA BHAT

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Doctoral Committee:

Adjunct Professor Richard Sproat, Chair
Professor Kenneth Church, Johns Hopkins University
Associate Professor Mark A. Hasegawa-Johnson
Professor Dan Roth
Professor Stephen E. Levinson

ABSTRACT

This dissertation is a study of two problems on estimation in the areas of natural language and speech. In the first problem we revisit the classical problem of estimating the size of unseen elements which we study in the context of a regime that is characterized by a large number of rare events, natural language being one. We propose an estimator of the size of the vocabulary of the underlying population that generates an observation and show that it has theoretical guarantees of optimal performance. Using natural language corpora from different languages we show that the performance of our estimator compares favorably with that of state–of–the–art estimators.

In the second problem, we explore the effect of vocabulary size and temporal aspects of speech production on perceptions of second language fluency with the aim of designing objective methods of fluency assessment from spontaneous speech. We show that articulation rate, phonation-time ratio, mean length of silent pauses and the number of silent pauses per second are aspects of speech production that are well correlated with human assigned scores of fluency. The measures of lexical use that we found to correlate well with fluency scores were the total number of words spoken (word tokens), the number of different words uttered (word types) and the number of words spoken once (*hapax legomena*). With the goal of objective fluency assessment without the use of automatic speech recognition, we show the utility of measures of temporal aspects of speech production that were obtained from direct signal-level measurements. Their use in a logistic regression framework for predicting fluency scores showed high agreement with scores assigned by human raters. An interesting experiment was exploring the difference in automatic assessment based on random snippets of the spoken utterance and that based on the complete utterance. Although the differences are not seen to be statistically significant at the 1% level, this opens avenues for further experimentation.

To my family, for their love and support

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to Prof. Richard Sproat for his guidance and support during my doctoral study. His patience with my non-linear mode of research was an encouragement. His penchant for interdisciplinary research was contagious and had a tremendous influence on my approach to problem formulation and solving.

My gratitude goes to Prof. Mark Hasegawa-Johnson, who contributed immensely to my research progress by way of insightful comments and interesting discussions.

My heartfelt thanks to Prof. Ken Church, whose support as a mentor and provider of valuable feedback on my dissertation is sincerely acknowledged.

I would like to acknowledge my friend Su-Youn Yoon's timely help in lending her second language spontaneous speech database to me. But for that, my experiments on fluency assessment would not have been possible.

Words fail me in acknowledging the extraordinary support of my loving family. Pramod, my husband, walked with me during all stages of my research, took on more than his share of the family responsibility, and is the person who made this degree possible for me. My loving daughter, at her young age, cooperated with me by putting up with my busy academic schedule and much to her dislike, had to do several activities of her interest without my participation. She understood that my study was as important as hers so I could bring in more ideas to share with her. My doting parents-in-law were ever ready to offer their support in ways they could and their ardent wish that I complete my doctoral degree has brought me so far. My parents and siblings have wished me well in all my endeavors.

To all fellow-students who made my stay here productive and enjoyable — Andrew Fister, Yoonsook Mo, Boon Pang Lim, Alla Rozovskaya, Lin Tan, Yuancheng Tu, Vinod Vydiswaran, Chen-Huei Wu, Su-Youn Yoon — Thank you!

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Estimation problems are a common feature in all engineering tasks. A canonical framework involves a set of observations used in estimating the value of a quantity of interest which has the appropriate engineering implication. The challenge is in the fact that the quantity being estimated and the observations are only *noisily* related. Classical and broad solutions address this setup. These are specifically relevant when the quantity being estimated is naturally related to the observations. In particular, there are two separate directions the solutions have taken:

- Statistical *signal processing* methods, where the topology of the space of observations is critically used. Examples include linear and nonlinear filtering (say, Kalman and Wiener filters) and kernel methods (say, support vector machines).

- *Bayesian* methods, where there is not much in the way of topology for the space of observations, but the likelihood of the observation given the target is obtained easily. Examples include hidden Markov models and naive Bayes classifiers.

Some problems in speech and natural language have specific features that render a routine application of classical solutions clumsy. In particular:

- When the statistical process generating the observations is very productive: natural language is a stark contrast to the language of digital communication. Specifically, the vocabulary size of natural language observations is very large, particularly when compared to the observation size (and contrasted with the binary observations standard in digital communication). This renders even very large sample sizes inadequate for making empirical frequency based estimates.

- When the quantity estimated is only *subjectively* related to the observations. For example, the observation in speech processing is an utterance (represented as a voltage waveform). Some typical quantities that one may be interested in estimating may be fluency of speech, speaker identity, language identity and pronunciation quality. Each is only loosely related to the utterance since there is no obvious class of functions relating the observed voltage waveform to the quantity being measured.

This dissertation concerns itself with two problems of estimation that belong to the class of problems just mentioned:

1. The main focus is on estimating the vocabulary size of the underlying population given a large sample. We consider a population that is characterized by a large number of elements with very small chance of occurrence in any given sample. This property renders even large samples of the population inadequate for inferring the underlying probability distribution over the elements. The problem of vocabulary size estimation is a classical one and our approach is to address the problem in a new regime, that of natural language. We propose an estimator of the number of unseen elements in a sample and show that it is statistically consistent. We then test its performance in the natural language domain by using corpora in different languages.

2. Our second focus is in the domain of language testing. Most human listeners perceive the level of fluency in speech quickly and reasonably uniformly. As such, fluency assessment has traditionally been done by human judges (example: standardized tests of spoken language). While there is a clear need to automate this process, there is no clear path: the main obstacle appears to be that the fluency level is *subjective*, i.e., there is no obvious function that relates the spoken utterance (physically a voltage waveform) to fluency level. We cast this as an estimation problem: the goal is to automatically estimate whether an utterance is fluent or not given a set of quantifiers that suitably measure perceptions of fluency. The engineering challenge is in the design and measurement of quantifiers that appropriately characterize aspects related to perceptions of human rated fluency and then in the design of an optimum estimator of fluency.

Here we would like to assess the effect of a person's vocabulary knowledge on the person's oral fluency. Additionally, we would like to investigate the extent to which a set of measures of temporal aspects of speech (known to correlate well with human perceptions of fluency) can be obtained via direct signal-level measurements. An important focus of our study is the design of alternate methods of automatic fluency assessment systems that are less reliant on automatic speech recognizers, which are known to perform poorly with spontaneous second language speech.

3. A third and minor focus is in the domain of web advertising where we consider the problem of predicting the probability of an on-line advertisement being clicked and propose a simple feature selection model. Query logs provide a large set of variables (example: IP address, day of the week) associated with the outcome (an ad being clicked or not). A key aspect of this setting is that the space of variables is very large compared to the size of training data. This aspect makes traditional estimation using empirical frequency measures moot. Our innovation in addressing this problem involved taking a two-step procedure and iterating it.

    - We first provide a simple algorithm that first reduces the space of the variables.
    - This reduced space of variables is next used in a (fairly standard) logistic regression model to estimate the click-through rate of advertisements.

    We iterate among these two steps using the feedback from the click-through rate performance at the end of each step.

In the next section we delineate the motivation behind the set of problems studied and unite the seemingly disparate set of problems by a common thread.

1.1 Motivation

1.1.1 Vocabulary Size Estimation

In several areas of study, the problem of estimating the number of classes in the underlying population, or equivalently, estimating vocabulary size, is of fundamental importance. For instance, questions such as *How many words did Shakespeare know?* or, *How many species of butterfly inhabit the Malayan region?* have been of great interest in classical studies [1, 2].

We first begin with the simple problem statement:

> Suppose that a random sample is drawn from a population. Further suppose that the population consists of identifiably different species. From the sample at hand, we can count the number of different species that have been seen and the frequencies with which they were observed. Using this information, can we say anything about the species that have not been seen in the sample? In particular, can we know the number of unseen species in the population?

In what follows we will describe this problem as it relates to areas as diverse as ecological studies, anthropology, engineering, genetics, corpus linguistics, psycholinguistics and language testing.

1. To biologists and ecologists who are concerned with assessing biodiversity and extinction rates of a population of plants or animals in a particular region, knowing the number of species of the population of interest is of fundamental importance.

2. In the field of anthropology, numismatists may be interested in assessing the monetary system of a given period in history. In such instances, estimating the number of coin dies and the number of coins per die provide the necessary information to conduct such a study [3]. Similarly, based on a sample of an extinct writing system it may be of interest to estimate the size of the alphabet of the writing system to make inferences about the writing system.

3. In studying system reliability it may be of interest to system engineers to know the number of errors in a software system.

4. In database management systems, a query optimizer determines the most efficient way to execute a query. Given an input query, the optimizer determines an efficient query execution plan (a set of steps used to access or modify information in an SQL relational database management system). One of the hardest tasks of a query optimizer is to determine the cost of a given query plan which relies heavily on the availability of statistics such as the number of unique values in a column. Thus accuracy of estimates of the distinct values in a column significantly impacts the query optimizer's ability to generate good execution plans for SQL queries. This in turn affords efficient processing of complex queries over large volumes of data [4, 5].

5. In designing speech recognition systems, one of the challenges is the out-of-vocabulary (OOV) rate. Given a training corpus, however large, there is always a set of words in the test corpus which are unseen in the training corpus and hence will not be recognized by the system. This set of words, called the out-of-vocabulary words, influences the accuracy of a system. Thus, the accuracy of a speech recognition system is no less than the out-of-vocabulary rate. Having an estimate of the number of out-of-vocabulary words gives an upper limit on the performance of a speech recognizer given its current training data and vocabulary. This is possible by estimating the vocabulary size of the fictitious population of which the training data is a sample.

   Having a large collection of documents for training data, an important criterion in the design of the training corpus for a speech recognizer is a vocabulary that will ensure good coverage of the training data and minimize the OOV rate.The vocabulary size can be made arbitrarily large, but this increase results in two conflicting effects: on the one hand, this results in a decrease in OOV related recognition errors; on the other, this increases the acoustic confusability caused by the introduction of more words, which in turn can be the source of recognition errors [6].There is also an associated cost of determining the accurate pronunciation for every vocabulary entry [7]. Having an estimate of the out-of-vocabulary words and hence the OOV rate can thus be used to guide the choice of an optimal vocabulary size and the to determine the usability of a training corpus for a speech recognizer.

6. In the area of information retrieval, an estimate of the vocabulary size of a large collection could be used to estimate the size and number of inverted lists that would be required to index it.

7. In genetics, estimating the total number of alleles (alternative DNA sequences) at a single locus in a population using the number of alleles observed in a sample is important for improving the characterization of the prior distribution for the allele frequencies, adjusting the estimates of genetic diversity, and estimating the range of microsatellite alleles [8].

8. In the areas of stylometry, authorship attribution and language acquisition, for instance, vocabulary size and the type-token ratio (the ratio of the distinct words to the total number of words used) are typically used to quantify vocabulary richness and lexical variety while analyzing and comparing corpora [9, 10]. Thus, one could ask to know the expected vocabulary size of the smaller corpus when the sample size is "stretched" to that of the larger one. Hence we need methods to extrapolate the seen vocabulary size or estimate the vocabulary size at an arbitrary sample size.

9. In the area of language testing, the measurement of vocabulary knowledge of second language learners is of interest to language teachers for assessing the students' language proficiency. Some situations in which estimates of vocabulary sizes of native and non-native speakers of a language are of interest are highlighted below [11, 12, 13].

    - An estimate of a person's receptive vocabulary size (the comprehension vocabulary used by a person in silent reading and listening) enables the teacher to assess the person's ability to deal with a range of language-related tasks that include reading and understanding texts such as newspapers, watching movies and engaging in friendly conversations.
    - Measures of a person's productive vocabulary (the vocabulary used in writing and speaking) enable assessment of writing quality.
    - Measuring vocabulary size is essential to chart the growth of learners' vocabulary to make comparisons across groups of non-native

speakers as well as for longitudinal assessment of an individual learner. This makes it possible to draw inferences about rates of growth in language learning. For instance, it can help answer questions such as whether non-native speakers increase their vocabulary knowledge at rates faster, slower or similar to those of native speakers.

- Estimates of native-speaker vocabulary sizes at different age levels provide moving targets for models of vocabulary acquisition for non-native speakers.
- Vocabulary size estimates are of interest to test developers who seek to develop reliable measures of second language ability.

Having seen that vocabulary size estimation is a fundamental question in several domains, we then seek to find the relation between vocabulary size, as represented by a language learner's mental lexicon, and that learner's language ability. In particular, we are interested in understanding the extent to which a speaker's lexical use influences his/her being perceived as fluent in the language. We hypothesize that an understanding of this problem, in combination with the results from previous studies on the effects of temporal aspects of speech production, will enable us to design suitable quantifiers of oral fluency. This will, in turn, aid the design of a system capable of scoring spontaneous speech for oral fluency automatically while emulating human performance as closely as possible. We will next motivate this goal of ours.

1.1.2 Automatic Fluency Assessment

The increasing need to perform in the global arena has brought about a corresponding need to learn a second language. Second language learners now have access to both human-assisted and computer-assisted means of learning. The computer-assisted resources available for language learners range from aspects related to grammar correction to pronunciation correction. Testing for language competence, spoken or written, however, is being done mainly by expert human judges. Expert human rating is indispensable for high-stakes testing purposes as in the case of testing foreign language competence of diplomats or language teachers. The potential for making language proficiency assessment widely available with minimum human intervention and

low associated expense motivates the move from expert-rated subjective language ability assessment methods to more objective non-rated methods. A typical scenario would be in computer-aided second language learning where the learner wishes to assess his level of fluency. Such objective assessments are made possible by automating the process of language proficiency testing.

Language proficiency is assessed with regard to the modalities of language ability — reading, writing, listening and speaking. These components of language ability can be broadly classified on the basis of how proficiency in the abilities is tested. Reading and listening abilities can only be tested indirectly (since it is impossible to measure comprehension as a process in the brain) by eliciting answers to a set of questions based on what is being read or listened to. Abilities in writing and speaking, on the other hand, are measured by seeing how well the test taker uses language for the intended communicative purpose in writing and speaking.

Several automated methods of assessing reading, writing and some aspects of listening ability have been developed in recent years. Automatic essay scoring applications have been proposed in [14, 15]. More recently, advances in natural language processing have enabled some aspects of listening ability to be automatically scored. Listening abilities that are currently being automatically scored are short answers eliciting factual responses [16] and forms of highly predictable speech tasks [17]. The predictable speech tasks that are being automatically assessed are speaker repetitions of prompted speech [17]. The scores generated by these applications correlate highly with those assigned by human raters, making these applications very reliable even for high-stakes testing scenarios.

Automatic testing of spontaneous speech is more challenging than that of other abilities. This is because, in addition to the challenges related to natural language processing (also found in, say, scoring essays), testing for spontaneous speech entails automatically recognizing non-native speech, which is a challenge in itself. What makes recognition of non-native speech hard is the pronunciation differences and erroneous language constructs of second language speakers that are drastically different from those of natives speakers. Current speech recognizers for non-native speech have recognition rates that are far from acceptable. For instance, a model that has been trained on native speech and tested on non-native speech [18] has a 50 percent word error rate, while Zechner et al. in [19] report a similar performance for a speech

recognizer trained on non-native data. In addition, owing to the highly unpredictable nature of spontaneous speech, pattern-matching approaches, that are routinely used in low-entropy speech such as read or prompted speech cannot be used for recognition.

Let us now see what it takes to build systems that assess language abilities automatically. Current paradigms for testing language abilities automatically are based on emulating the human scoring mechanism. Accordingly, the process of engineering automated systems broadly follows these steps:

1. A set of criteria that influence the subjective decisions involved in human assessment are identified.

2. Suitable quantifiers serving as objective approximations of the criteria are chosen.

3. Algorithms to measure the quantitative variables and to combine them appropriately to approximate the human scoring process are designed.

At the identification stage, a set of criteria that most influence the subjective decision are listed and the human scorers are asked to carry out the assessment process on a large enough data set following the criteria very closely. For instance, in essay scoring the typical criteria of evaluation include grammatical accuracy, lexical choice, topical coherence and development [14]. Having chosen a set of criteria, the next step is to choose a set of objective measures that represent the criteria as closely as possible and also have a good coverage of the identified criteria. Taking examples from essay scoring again, measures of word frequency are used for measuring lexical complexity and topic-related word usage. Finally, methods of measuring the proposed set of quantifiers and combining them appropriately are explored. Engineering techniques that work at the heart of these systems leverage statistical techniques to handle the approximation process towards automating the assessment. Accordingly, the quantifiers are then combined using statistical models such as classification and regression trees or multiple regression models to produce automatic scores.

While the area of testing for predictable language abilities such as reading and essay writing has shown considerable maturity, research on automated methods of making assessments on spontaneous speech, however, is still in its infancy. The state-of-the-art system [19] for scoring second language

spontaneous speech currently uses automatic speech recognizers to assess spontaneous speech with limited success. It is operational serving as a personal evaluation tool for online test takers of the Test of English as a Foreign Language (TOEFL) since 2006, but is mainly used for data collection. Continuing explorations in this area, our study concerns itself with automatic assessment of spontaneous speech by means of signal-level acoustic measurements. Of particular interest to us is the task of automatic assessment of oral fluency in second language spontaneous speech.

Oral fluency is an important feature of speech which is considered a benchmark of evaluation of a person's proficiency in a language. Theories of language proficiency regard oral fluency as "low-level proficiency," an essential component of overall proficiency [20]. Apart from the general connotation of proficiency in a language, the notion of fluency has no agreed definition. While the definition of fluency varies among expert human raters, human ratings seem to reflect a tacit agreement on the notion of fluency. Fluency in a second language seems to be centered around two core components:

- the speaker's ability to speak *effortlessly* and *quickly*;
- the speaker's ability to communicate *effectively* — to be able to get his/her ideas across despite problems with the grammar, pronunciation and vocabulary.

Clearly, this indicates that central to understanding the notion of fluency is the ability to develop a multidimensional assessment of the effort of production and effectiveness of communication, command of pronunciation, grammar and vocabulary. As a step towards developing non-rated (automated) methods of language assessment, we need to be able to quantify the criteria judged by human raters. This study is an effort to develop automated methods of assessing language fluency as perceived by human raters. Our study aims to quantify perceptions of fluency along the components of

1. effort of production,
2. effectiveness of speech.

In particular, we seek to study the extent to which perceptions of fluency are influenced by

- temporal aspects of speech quantifying the speaker's effort of production, which we call the *quantitative* aspect of speech, and
- lexical aspects quantifying vocabulary use and in turn indicating effectiveness or the *qualitative* aspect of speech.

A special feature of our study is the use of methods that are not based on automatic speech recognition. We intend to explore the use of signal-level measurements as quantifiers of fluency.

While the first two problems deal with vocabulary size estimation and assessing effects of vocabulary use, respectively, the second problem also deals with approximating a subjective quantity, that of human perceptions of fluency, by means of a set of objective measures of fluency scores. The third problem that we consider in this thesis is also one of approximating a subjective decision, that of a user clicking on a web advertisement, by means of a set of user session related measurements. In the following subsection we will briefly consider motivations behind solving this problem.

1.1.3 Click-Through-Rate Prediction for Advertisements

Advertisements play an important role in search engine dynamics and economics. Choosing the appropriate ad for a query and the order in which it is displayed determines the chance that the user clicks on the ad. Thus, being able to accurately predict whether a given advertisement will be clicked or not greatly affects user experience. At the same time it strongly influences search engine revenues. Hence, accurate prediction of the click-throughrate of an advertisement given a set of variables that correspond to the ad is a fundamental problem.

We consider the problem of predicting the probability of a click for an advertisement when the outcome of a click or no-click is expressed by means of a set of variables. The 42 variables represent outcomes such as the following:

- query related quantities such as `MatchedKeyword` which stands for the number of words in the query that match with the keywords associated with the ad;
- advertisement related quantities such as `ListingID`;

- user related quantities such as `IP address` and `Age`;and
- some general quantities such as `DayOfWeek`.

The *outcome* associated with this set of measurements is a one or a zero indicating whether the advertisement was clicked or not. The data was obtained from Microsoft's proprietary query logs over a period of several months.

The problem of estimating the probability of a click given a set of variables can be viewed as one of estimating the conditional expectation of the outcome given the values taken by the associated set of variables. If the variables were very few, then we could empirically learn the *joint* statistics between them and the single outcome. This could then be used to design an algorithm that predicts the outcome given the variables. Here, however, the number of variables is very large. Thus there are issues of data sparsity in the observed data. Some variables take values in a large set of choices many of which may not have been observed in the sample. Even those that have been observed may only have occurred a small number of times. This makes the task of finding a good estimate of the joint statistics from the training data hard.

However, we have empirical estimates of the joint statistics between the outcome and each of the variables from the observed data. Our approach is to extrapolate these estimates to obtain the necessary conditional expectation of the outcome given the variables.

Variable selection is an important problem in the design of predictors in areas of application for which data with a large number of variables are available. This is the task of selecting a subset of variables that is most useful in building a good predictor. The objective of variable selection is to:

- improve the prediction performance by avoiding over-fitting;
- build cost-effective and faster predictors by reducing data collection and storage costs and reducing computational effort;
- provide better understanding of the underlying process that generates the data;
- provide some immunity from the possibility of missing values.

Hence there is a need to rethink approaches that design predictors using all the variables and study alternative methods that employ effective variable

selection strategies. This motivates our considering a suitable approach to variable selection.

The problem of variable and feature selection in domains with several hundreds of variables has been addressed in several works as presented in the surveys [21] and [22]. Other works [23] have shown improved accuracy in classification tasks when the data is represented using a reduced number of relevant features. The present work investigates the applicability of feature selection in the area of web advertising. More specifically, we study variable selection in the task of predicting the click-through rate of an advertisement when the outcome (click or no-click) is expressed by means of a large set of variables.

In summary, this study aims at answering three research questions belonging to the two broad estimation classes mentioned before:

- We are interested in the theoretical problem of modeling a regime characterized by a large number of rare events with low probabilities of occurrence and deriving an estimator of vocabulary size in such a regime with performance guarantees.

- We are interested in estimating a target quantity, quantified by a set of domain specific measures, by approximating the target quantity in terms of its quantifiers. We study this problem in two engineering scenarios:

    1. Design of an automatic oral fluency assessment system for rating second language spontaneous speech. Toward this end, we first assess the effects of quality of speech production (measured via lexical use as indicative of the person's productive vocabulary) and quantity of speech production (measured using temporal variables such as speech rate) on human perceptions of oral fluency and obtain a set of quantifiers of fluency. We then approximate the human assigned fluency score by means of this set of quantifiers.

    2. Design of a click-through-rate prediction system for on-line advertisements by choosing an optimal set from among a large set of quantifiers of the probability of click-through. The probability of click-through is then approximated by a combination of the quantifiers in the optimum set.

1.2 Thesis Overview

This thesis is organized in the following manner:

In Chapter 2, we survey of the literature available on the problems addressed in this study. We consider the different approaches to vocabulary size estimation that have been proposed in several domains and get a feel for the state of the art. In the context of automatic assessment of fluency, we look at the various measures that have been studied before. Subsequently, we take a look at the state-of-the-art system—the *Speechrater*, one component of which is a fluency assessment module.

In Chapter 3, we take a renewed look at the classical problem of vocabulary estimation in a regime characterized by the presence of a large number of rare events. We point out a short-coming of an estimator that is based on the Good-Turing estimator of probability mass of unseen events. We then propose our estimator to address the short-coming and show its statistical consistency. Subsequently, we evaluate the performance of our estimator in comparison with other state-of-the-art vocabulary size estimators by using natural language corpora in different languages.

In Chapter 4, we study the effect of vocabulary size and temporal aspects on speech production on human perceptions of fluency. We use the results of this study to design an automatic system that rates whether a given spontaneous utterance is *fluent* or *not fluent*. Additionally, we study thin-slice based judgment of fluency where we compare automatic assessment based on a random snippet of the utterance with that based on the entire utterance.

In Chapter 5, we study the problem of estimating the probability of a click of an on-line advertisement when the outcome of being clicked or not is represented by means of a large set of quantifiers, many of them weakly related to the outcome. We propose an incremental feature-selection model that chooses the best set of features to predict the probability of a click.

In Chapter 6, we summarize and interpret our results and provide directions for possible future work.

CHAPTER 2

BACKGROUND AND RELATED WORK

In this chapter we will survey existing literature on the problems we study in this thesis. We first present previous work on vocabulary size estimation, following which, we consider previous work on quantifiers of fluency and automatic language ability assessment systems.

2.1 Vocabulary Size Estimation

In the previous chapter we saw some domains where vocabulary size estimation is of fundamental importance. We will now consider the estimators that have been proposed to address the problem. We first review the available estimators from the works in statistics and then proceed to those estimators proposed in the area of corpus linguistics and finally to those proposed in the domain of language testing. The estimators differ fundamentally in their modeling approach; with the exception of the domain of language testing, all the estimators are based on probabilistic formulations. In some cases, the sample is considered to be drawn according to an underlying probability law, such as the Poisson law, or the multinomial distribution. In other cases, the population frequencies of the individual classes are assumed to have some parametric form, such as the Zipf law for word frequencies. In language testing, estimates of vocabulary size are mainly dictionary-based and are in turn functions of the accuracy with which words from a sample of a dictionary are recognized.

2.1.1 Terminology and Notation

A sample of size n is drawn from a population with C classes, where C is unknown. In what follows, the total number of classes C will equivalently be

termed as *vocabulary* Ω and the individual classes as *words*. A sample from the population, represented by the observation vector

$$(X_1, \ldots, X_n)\,, \tag{2.1}$$

from a large, but finite, vocabulary Ω. In theory we have the random vector $\mathbf{n} = (n_1, n_2, \ldots, n_C)$, where n_i, for $i = 1, \ldots, C$, is the number of observations from class i in the sample. The ith class appears in the sample only if $n_i > 0$, but we do not know from the sample which of the n_i's is zero. This means that the vector $\mathbf{n}$ is not observable. What we observe from the sample, however, is the vector $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_n)$, φ_k being the number of classes represented k times in the sample. Thus, φ_1 is the number of singletons, φ_2 is the number of classes observed two times, and so on, in the sample. We will call φ the *spectrum* of the sample and φ_k the individual *spectrum element* corresponding to frequency k (also termed as *frequency of frequency* k in [24] and denoted by N_k in sources such as [25]). The problem, then, is to estimate C based only on the observable, the spectrum. We denote by V the number of classes in the sample (or the seen vocabulary), so that $\sum_{k=1}^{n} k\varphi_k = n$, and $V = \sum_{k=1}^{n} \varphi_k$. We denote by $\hat{V}$ the estimate of the number of classes, C in the population.

2.1.2 Statistical Estimators

Standard reviews of the works on estimating the number of species in a population are [26] and [27]. The estimators that have been proposed have been prompted by domain specific problems. For instance, several of the estimators available have been proposed in the context of population size estimation of plants and animals. Others have been proposed to estimate the size of an author's vocabulary. Several others have been proposed in applications as diverse as estimating the number of distinct records in a filing system where many records are duplicated, undiscovered observational phenomena in astronomy and errors in a software system. We resort to the survey of the statistical models and the taxonomy of the estimators as found in [26] and represented in Figure 2.1 .

Resorting to the notation in Section 2.1.1, our goal is to estimate the "essential" size of the vocabulary Ω using only the observations. In other

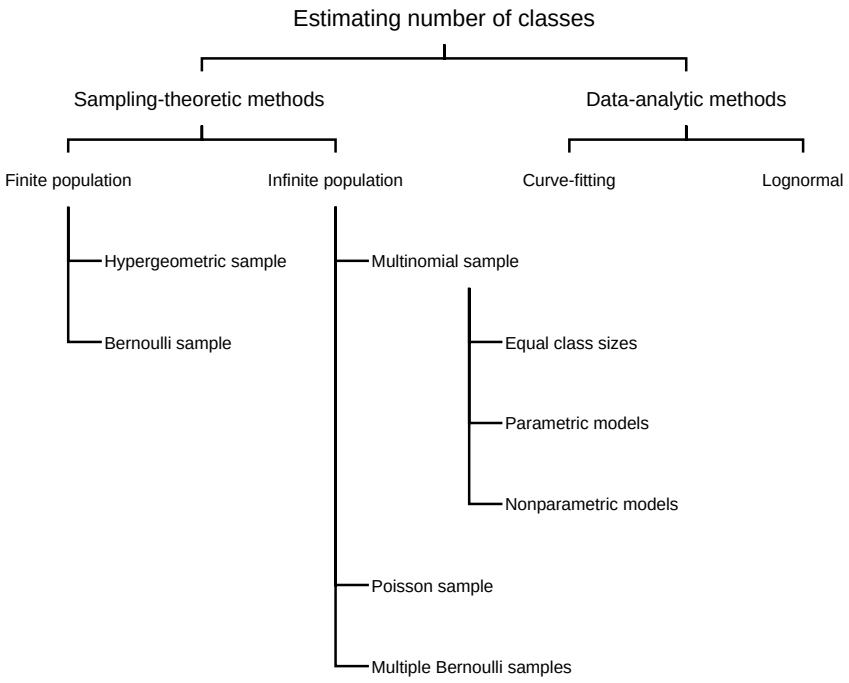

Figure 2.1: Classification of the statistical estimators of vocabulary size in existing literature.

words, having seen a sample of size n we wish to know, given another sample from the same population, how many unseen elements we would expect to see.

Estimators

We consider the estimators following the tree diagram.

1. Sampling-theoretic methods: Estimators in this category consider a sample to behave according to a probability law.

   (a) Finite population

   Hypergeometric sample:

   Suppose the population is finite with known size N and let N_i denote the number of units in the ith class, $i = 1, \ldots, C$. This can occur, for instance, in sampling a database for duplicate records. If we draw a sample at random without replacement from this population, then **n** has a hypergeometric distribution. For such a model, Goodman [28] showed that when the sample size is known to have the property $n \geq max_{1 \leq i \leq C} N_i$, an unbiased estimator of C is

   $$\hat{V}_{Goodman1} = V + \sum_{k=1}^{n} (-1)^{k+1} \frac{(N-n+k-1)!\,(n-k)!}{(N-n-1)!n!} \varphi_k.$$

   It was found that the performance of the estimator depended critically on the sample size. Taking an asymptotic approach, Shlosser [29] showed that as N and n are made arbitrarily large such that n/N is a constant $q \in (0, 1)$, an estimate of the number of classes is given by

   $$\hat{V}_{\text{Shlosser}} = V + \varphi_1 \left(\sum_{k=1}^{n} kq\,(1-q)^{k-1} \right)^{-1} \sum_{k=1}^{n} (1-q)^k \varphi_k.$$

   The performance of this estimator was shown to be better than that of $\hat{V}_{\text{Goodman1}}$ even for small sample sizes.

   Bernoulli sample:

Suppose that the N items of the population are drawn in the sample independently, each with probability p. Then the sample can be considered to be binomially distributed with parameters (N, p). This model has been used by numismatists, for instance, for the appearance of coins in a collection. We mention two of the estimators [28, 30] that have been proposed for this model. The first one is

$$\hat{V}_{\text{Goodman2}} = V + \sum_{k=1}^{n} (-1)^{k+1} \left(\frac{1-p}{p}\right)^{k} \varphi_k.$$

This estimator was also found to be critically dependent on the size of the sample relative to that of the population and as such was considered unusable. The second one considered here, due to Esty [30], was obtained by considering the Bernoulli sample as a sample of the *superpopulation* which has a negative binomial distribution. The estimator is

$$\hat{V}_{NB} = \frac{n}{\hat{\mu}},$$

where $\hat{\mu}$ is a function of the parameters of the negative binomial distribution of which the population is considered a sample. The estimator, however, was found to be unacceptable based on simulation results [31].

(b) Infinite population

We next consider models when the population size is infinite. We thus have a random sample of n items from an infinite population that is partitioned into C classes with probabilities $\pi = (\pi_1, \ldots, \pi_C)$, such that $\sum_{i=1}^{C} \pi_i = 1$. The modeling proceeds by making a finite population approximation.

Multinomial sample, equal class sizes:

Here, the classes are assumed to be equiprobable (probability of a given class i is $\pi_i = C^{-1}$) in the population to make the model tractable. This case has a vast literature [32], and constitutes part of the classical *occupancy* and *coupon collector's* problems addressed in the associated literature.

The maximum likelihood estimator $\hat{V}_{\text{MLE}}$, of the vocabulary size, is then given by the solution $V^\star$ of the equation

$$V = V^\star \left(1 - e^{-n/V^\star}\right),$$

as shown in [33]. That $\hat{V}_{MLE}$ underestimates the number of classes when the classes in the population are not equally likely, has been pointed out in [34].

In this context of equiprobable class assumption, we include a coverage-based estimator as found in the literature [26, 27]. The coverage, u, of a sample is the (random) sum of the population probabilities (π_i's) corresponding to the observed classes. Thus, $u = \sum_{i=1}^{C} 1(n_i > 0)\pi_i$, where $1(A)$ is the indicator function of the event A. When the classes are equiprobable, $u = V/C$; so given an estimator of $\hat{u}$ of u, an estimate of the number of classes is $V/\hat{u}$. The first such $\hat{u}$ proposed by Good in [24] was $\hat{u}_{GT} = (1 - \varphi_1/n)$ and its utility even when the equiprobable class assumption is not valid has been established in [35] . Under the assumption of equiprobable classes, then,

$$\hat{V}_{\text{Cov}} = \frac{V}{\hat{u}_{GT}}. \tag{2.2}$$

That this estimator (termed as the Good-Turing estimator of vocabulary size) is best suited for distributions where the equiprobable assumption holds good has been empirically observed in [27]. Owing to its desirable asymptotic properties and ease of computation, it has been stated [26] that it may be preferable in applications.

Multinomial sample and Parametric models:

In applications where the classes are not equally likely, but where some classes have very low chances of occurring, estimators of the number of classes have been proposed by resorting to parametric modeling techniques of two types: in the models of the first type, the population relative frequencies of the classes, π_i's, are assumed to have a functional form that depends on a set of parameters; in the models of the second type, the histogram of the π_i's are

approximated by a probability density function that depends on a set of parameters. Such a modeling affords the use of population probabilities or a probability density function of the population probabilities for estimating the spectrum elements, which we then use to estimate the number of classes in the population.

We will now consider the two types of models:

i. Parametric models of the population probabilities: The functional forms that have been considered in the literature (such as in [36]) are

   A. the Zipf model, where, $\pi_i = f(i;\theta,C) = \theta/i$, $i = 1,\ldots,C$, and

   B. the Mandelbrot model, where $\pi_i = f(i;\theta,C) = \theta_1/\left(\theta_2+i\right)_3^{\theta}$, for $i = 1,\ldots,C$.

   The parameter vector θ of the functional forms is estimated using the observed sample. The originally proposed estimators are involved and will not be considered here. However, Baayen in [37] and Evert in [38] consider computationally tractable approximations to the estimators which will be considered in the context of discussing estimators proposed in the domain of corpus linguistics (Section 2.1.3).

ii. Parametric models for the probability density function of the population probabilities: In these models, the histogram of the population relative frequencies is approximated by a probability density function that depends on a set of parameters. Sichel [26] modeled the probability density function of the population probabilities as having a generalized inverse Gaussian distribution $\psi\left(\pi;\theta_1,\theta_2\right)$ of two parameters θ_1 and θ_2. Under this assumption, he then derived an estimate $(\hat{\theta}_1,\hat{\theta}_2)$ of (θ_1,θ_2) which was based on the observed spectral elements. He then used this to obtain an estimate of the number of classes as

$$\hat{V}_{\text{Sichel}} = \frac{2}{\hat{\theta}_1\hat{\theta}_2}.$$

Of the two types of parametric models it has been stated that the second seems preferable for applications owing to its more

general nature in that it seeks to specify not the exact population probabilities, but rather their distribution [26].

Multinomial sample, nonparametric models:

The models discussed thus far made certain assumptions about the population probability vector π with the probabilities being all equal, or having an underlying probabilistic law (parametric models). Nonparametric models have been proposed for estimating the number of classes in the population *without* making assumptions on π. This would mean that the estimator would then have to be based solely on the spectrum elements and assumptions thereof. Bunge and Fitzpatrick in [26] mention works that showed that no unbiased estimate exists and that the bias of any estimator of the number of classes C based on the frequency spectrum is unbounded over the set of possible populations. Nevertheless, Chao in [39] used estimates of the moments of the spectral elements to obtain a nonparametric estimator,

$$\hat{V}_{\text{Chao1}} = V + \frac{\varphi_1^2}{2\varphi_2}. \tag{2.3}$$

Another estimator proposed by Chao and Lee [40]used the idea of coverage of a sample (and the popular estimator of coverage due to Good [24] given by $\hat{u} = 1 - \varphi_1/n$) to derive the nonparametric estimator

$$\hat{V}_{\text{Chao2}} = \frac{V}{\hat{u}} + \frac{n\,(1-\hat{u})}{\hat{u}}\hat{\gamma}^2, \tag{2.4}$$

where $\hat{\gamma}$ is the estimate of γ, the interspecies variance, given by

$$max\left(0, \frac{nN}{n-\varphi_1}\frac{\sum_{i=1}^{n} i\,(i-1)\,\varphi_i}{n\,(n-1)-1}\right). \tag{2.5}$$

A related nonparametric estimator has been proposed by Gandolfi and Sastri in [27]. The estimator is Bayesian and is given by

$$\hat{V}_{\text{GS}} = \frac{nV}{n-\varphi_1} + \frac{n\varphi_1}{n-\varphi_1}\gamma^2,$$

where

$$0 \leq \gamma^2 = \frac{-n - V + \varphi_1 + \sqrt{5n^2 + 2n(V - 3\varphi_1) + (V - \varphi_1)^2}}{2n} \leq 1.$$

By means of simulations on different data sets, this estimator has been empirically found to be more accurate than the available non-parametric estimators on data sets with nonuniform population probabilities while performing poorly in the uniformly distributed data sets.

Poisson sample:

Now suppose that the samples consists of representatives from different classes of the population. Further suppose that the number of representatives of the ith class in the sample is a Poisson random variable with mean $\lambda_i, i = 1, \ldots, C$, and that these variables are independent. This model was proposed by Fisher in [2]. If we assume that the λ_i's are themselves a random sample from some distribution F, then $E(V) = C(1 - p_0(F))$, where $p_0(F)$ is the probability that an F-mixed Poisson random variable is equal to zero. Thus, given an estimate $\widehat{p_0(F)}$ of $p_0(F)$, an estimator of C is

$$\hat{V} = \frac{V}{1 - \widehat{p_0(F)}}. \tag{2.6}$$

Efron and Thisted in [1] use this model to estimate the number of words that Shakespeare knew and did not use.

Multiple Bernoulli sample:

Suppose that an infinite population (partitioned into C classes) is observed on each of n occasions, and on each occasion each class either is or is not observed. The sample, consisting of observations over the n occasions, can then be represented by the $C \times n$ matrix $[x_{ij}]$, where $x_{ij} = 1$ when the ith class is observed on the jth occasion, $i = 1, \ldots, C, j = 1, \ldots, n$. Burnham and Overton in [41] studied such a model in which the x_{ij}'s are all independent and $P(x_{ij} = 1) = \pi_i$ for $i = 1, \ldots, n$ (the probability of observing a class is the same on each occasion). Taking the π_i's to be a random sample from some distribution F, they developed a kth order jackknife estimator. An instance of this is the first order

jackknife estimator given by

$$\hat{V}_{BO1} = V + \left(\frac{n-1}{n}\right)\varphi_1.$$

The estimation performance of this class of estimators, according to Bunge and Fitzpatrick, is unclear since studies regarding their performance have not been conclusive.

2. Data-analytic methods: The main data-analytic methods considered in [26] are, the method of extrapolation of curves and the lognormal fit, which we will briefly cover next.

   Extrapolation of curves:

   For many of the problems discussed above, one can in principle derive a graph of the expected number of observed classes as a function of the sample size n, denoted by $E(V^{(n)})$. Depending on the particular problem, this can be a *coupon collector's*, *type-token* or *species-area* curve, obtained from the observed values of the spectrum for several sample sizes. Using the functional form of the curve and the observed spectral elements, it may be possible to estimate the total number of classes by extrapolation, without reference to a sample-theoritic model. In other words, suppose that we assume only that:

   (a) $E(V^{(x)}) = f(x;\theta)$, where x is some measure of the size of the sample (not necessarily the count of items), θ is a parameter vector, and f is a given increasing function of x; and,

   (b) $lim_{x\to\infty} f(x;\theta) = C$.

   Under these assumptions, if an estimate $\hat{\theta}$ of θ can be obtained from the spectral elements, then the required estimate $\hat{V}$ will be $lim_{x\to\infty} f(x;\hat{\theta})$. According to Bunge and Fitzpatrick [26], although a few estimators based on this method have been found in the literature, "it is hard to be very optimistic about the potential of such methods, because if the function $f(x;\theta)$ is derived from the sampling model, the sampling theory will give a more efficient estimate of C, and if it is not, then its form seems difficult to justify."

   Lognormal fit: Preston [42] found that the graph of the spectral

elements (φ_k) versus logarithm of the spectral unit ($log_2 k$) often resembles a Gaussian curve, truncated on the left. An estimate of C can be obtained by fitting a Gaussian curve to the curve based on the spectrum elements, extrapolating it to the left, and integrating it over $(-\infty, +\infty)$. Very few studies have resorted to this approach, and these have seen little success.

In discussing the state of the art after comparing several estimators of the number of classes, Bunge and Fitzpatrick point out that "it is rare that a sampling model obtains exactly; ideally one would like to have an estimator based solely on φ, the frequency spectrum, that is robust across various sampling plans and population structures." They then go on to state that without precise knowledge of the population and the sampling plan, the recommended estimator would be $\hat{V}_{\text{Chao2}}$, which is expected to be robust against deviation from the sampling plan as well as account for cases where the classes are not equiprobable.

2.1.3 Vocabulary Size Estimators in Corpus Linguistics

Literature on corpus linguistics mentions that the state-of-the-art methods of predicting vocabulary size and the number of singletons for different sample sizes are based on the statistical models of word frequency distributions [37]. Following the taxonomy of the previous section, the estimators are parametric versions of the infinite population, multinomial case. We have seen that this entails having parametric models of the population relative frequencies or histograms of relative frequencies which are in turn used to derive estimators of spectral elements (and hence vocabulary size) as a function of sample size. Evert and Baroni [38] carry out an evaluation of the extrapolation quality of these models to which we turn next.

Zipf's law [43] stipulates that the probability π_i of a word w_i is inversely proportional to its Zipf rank r_i, the rank of the word w_i in the list of all words ordered by decreasing frequency. In its most general form, the law is given by

$$\pi_i = C r_i^{-\alpha}, \tag{2.7}$$

where $1 < \alpha < 2$ and C is a normalizing constant to ensure that π_i's are probabilities. An extension to this law, called the Zipf-Mandelbrot law (ZM

law), attempts to account for the lack of fit of the Zipf's law in the low frequency region. Its functional form is given by

$$\pi_i = \frac{C}{(r_i + b)^a}, \tag{2.8}$$

where $a > 1$ and $b > 0$. Using this formulation, Evert in [44] arrives at closed form expressions for expected values of the spectral elements φ_k and vocabulary size as functions of sample size. Accordingly we have

$$\hat{\varphi_k} = \frac{C}{k!} n^\alpha \Gamma(k - \alpha), \tag{2.9}$$

and

$$\hat{V_{ZM}} = C n^\alpha \frac{\Gamma(1 - \alpha)}{\alpha}, \tag{2.10}$$

where $C = \frac{1-\alpha}{B^{1-\alpha}}$ and $\alpha = 1/a$ and $B = \frac{(1-\alpha)}{b\alpha}$, a and b being the constants of the Zipf-Mandelbrot fit to the sample. Evert notes that the appeal of the ZM model lies in its mathematical elegance and numerical efficiency.

While the ZM model considered above assumes an infinite vocabulary, for a finite vocabulary scenario, Evert obtained a related ZM formulation, the finite Zipf-Mandelbrot model, for obtaining the expected vocabulary size and spectral elements as functions of sample size. In this model, they are calculated to be

$$\hat{\varphi_k} = \frac{C}{k!} n^\alpha \Gamma(k - \alpha, nA), \tag{2.11}$$

and

$$\hat{V_{fZM}} = C n^\alpha \frac{\Gamma(1 - \alpha, nA)}{\alpha} + \frac{C}{\alpha A^\alpha}(1 - e^{-nA}). \tag{2.12}$$

2.1.4 Nonstatistical Estimators

The estimators that we consider in this category are mainly those of receptive and productive vocabulary sizes. These are routinely used in language testing where the goal is to estimate the size of the mental lexicon of the subject being considered. A typical application scenario is in assessing the vocabulary size of a second language learner for the purpose of judging the person's language proficiency as is done in the Test of English as a Foreign Language (TOEFL).

Vocabulary size tests in English estimate a learner's vocabulary size using a graded sample of words covering numerous frequency levels [13]. The

words are units identified as being relevant for the purpose of testing and are elements of sets called word families, which consist of a base word together with its inflected and derived forms that share the same meaning. For example, the word forms *extends, extending, extended, extensive, extensively, extension* and *extent* are all members of a word family headed by the base form *extend*. The frequency levels themselves are defined by reference to word-frequency data which are obtained from standard English corpora. The learner's knowledge of the sampled words are tested and the results give an approximation of the proportion of the total number of words at each frequency level that the learner knows.

2.1.5 Summary of Estimators of Vocabulary Size

We saw that the statistical estimators of vocabulary size have different modeling assumptions and that the applicability of an estimator to a particular problem calls for a thorough assessment of the assumptions underlying the design of the estimator as well as of the domain under study. However, there are a few that stand out as being nonparametric, and hence have been ascribed a wider appeal in terms of application [26]. Notable among them are the estimators by Chao-Lee, V_{Chao1}, V_{Chao2} and the coverage-based estimator V_{Cov} (also called the Good-Turing vocabulary estimator V_{GT} [27]). When the underlying distribution is nonuniform (the classes are not equiprobable), the nonparametric estimator V_{GS} has been found to perform well.

In the area of corpus linguistics we saw that the estimator based on the Zipf-Mandelbrot model for observed frequencies and its version for the finite vocabulary case are state of the art for extrapolating vocabulary sizes to larger samples based on an observed corpus. In the area of language testing we saw that the assessment of vocabulary size is based on identifying words representing different frequency and predictability levels from the language.

In Chapter 3 we propose an estimator of vocabulary size and discuss its statistical properties. Subsequently, we compare its performance alongside that of the state of the art estimators mentioned above.

2.1.6 Probability Estimation of Rare Events

Probability estimation of rare events is an old problem. Laplace (1825) considered the probability that the sun may not rise tomorrow. In their effort to break the German Enigma code in WWII, Good and Turing (1941) worked to find a pattern in the passwords used by the German U-boat commanders. A series of recent works by Orlitsky et al. [45, 46, 47, 48, 49] has shed new insight into these classical probability estimators. They provide a common framework to study the estimators and some simple modifications that they show to vastly improve the performance of these estimators. In this section, we review this problem and summarize the main results of Orlitsky et al. Further, the results of Orlitsky et al. suggest a natural solution to the problem of interest in this thesis: estimation of the number of unseen elements in a data set. We conclude this section by studying this alternative viewpoint; we show that the utility of this alternative approach depends critically on the solution to what appears to be a computationally very hard problem.

Probability Estimation

A classical scenario that is at the heart of many engineering problems is that of estimating the probability distribution of a sequence of observations. A simple mathematical model states it as follows: let $X_1, \ldots, X_n$ be a sequence of i.i.d. (independent and identically distributed) random variables derived from an underlying probability distribution $\mathbb{F}$ over a vocabulary Ω. The classical problems ask for an estimation of the underlying distribution $\mathbb{F}$.

The assumptions made on the vocabulary size ($|\Omega|$) greatly impact the type of estimators and what is known about this problem. In particular, we can consider two (very different) regimes:

- Large sample size and small vocabulary, and
- Small sample size and large vocabulary.

Large Sample Size, Small Vocabulary This is a classical setup: the theoretical understanding is quite deep and is well implemented in practice as well. In particular, the *empirical* estimate

$$\hat{\mathbb{F}}_n(x) := \frac{\sum_{i=1}^{n} \mathbf{1}_{X_i = x}}{n}, \quad x \in \Omega,$$

is the *maximum likelihood* (ML) estimate of the true underlying distribution $\mathbb{F}$. Here we have used the notation

$$1_{X_i=x} = \begin{cases} 1 & \text{if } X_i = x \\ 0 & \text{else.} \end{cases}$$

It is known to be *consistent*, i.e.,

$$\hat{\mathbb{F}}_n(x) \to \mathbb{F}(x), \quad x \in \Omega,$$

as $n \to \infty$.

This setup occurs in many practical situations, the foremost of which are reliable communication over noisy channels and lossless data compression (of, say, ASCII or binary files where the vocabulary is small). The maximum likelihood estimate is used (at least in a motivational sense) in practical communication and data compression schemes (parity check codes and programs such as `gzip` and `compress`).

Small Sample Size, Large Vocabulary The setup with small values of n, as compared to the $|\Omega|$, is much less well studied in the literature. But it shows up in several engineering problems naturally: language modeling, compression of natural languages, (lossy) compression of images and video, to mention a few. The ML estimate can be *wildly* inappropriate for this scenario, the trouble being that many words of the underlying vocabulary are never observed in the data set, but they still have significant probability of occurring. This issue is best explained in the words of Orlitsky et al. [46]:

> In preparation for your next safari, you observe a random sample of African animals. You find 3 giraffes, 1 zebra, and 2 elephants. How would you estimate the probability distribution of the various species you may encounter on your trip?
>
> A "naive" empirical-frequency estimator will assign probabilities

$$\begin{array}{ll} \text{Zebra} & \frac{1}{6} = 0.17 \\ \text{Elephant} & \frac{1}{3} = 0.33 \\ \text{Giraffe} & \frac{1}{2} = 0.5 \\ \text{Other Animals} & \frac{0}{6} = 0. \end{array} \tag{2.13}$$

> But this estimate is clearly amiss, as the poor estimator will be

completely unprepared for an encounter with an offended lion.

Laplacian Estimator

This classical problem has been around at least since the time of Laplace, who proposed the famous "add one" estimator to estimate the underlying probability distribution by adding one to every observed element in the sequence [50]. Laplace proposed his estimator in the context of estimating the chance that the sun may not rise tomorrow. If we have observed the sun rising the last n days, the chance, according to Laplace, that it may not rise tomorrow is $1/(n+2)$. If Laplace had been on the safari above, he would have assigned the following probabilities:

$$\begin{array}{ll} \text{Zebra} & \frac{1+1}{10} = 0.2 \\ \text{Elephant} & \frac{2+1}{10} = 0.3 \\ \text{Giraffe} & \frac{3+1}{10} = 0.4 \\ \text{Other Animals} & \frac{0+1}{10} = 0.1. \end{array} \tag{2.14}$$

Good-Turing Estimator

Somewhat more recently (at least, relative to Laplace), the Good-Turing estimator was proposed during WWII (in the context of cracking the German `enigma` code) [51]. There are several variants of the Good-Turing estimator, but the basic version is the one introduced in Section 3.2, in the context of the total vocabulary estimation problem. Below is a statement of the Good-Turing estimator [52].

Given the observation $\bar{x} \stackrel{\text{def}}{=} (X_1, \ldots, X_n)$, we have the corresponding sequence of nonnegative integers $\varphi_1, ..., \varphi_n$, where φ_k denotes the number of distinct symbols that appear exactly k times in $\bar{x}$, for each $k = 1, \ldots, n$. In other words, φ_1 is the number of *singletons*, φ_2 is the number of *doubletons*, and so forth, in the sequence $\bar{x}$. The Good-Turing estimator assigns equal probability to all symbols $x \in \Omega$ that appear the same number of times. Specifically, a probability of

$$\frac{(k+1)\varphi'_{k+1}}{n\varphi'_k}, \quad k = 1 \ldots n, \tag{2.15}$$

is assigned to all symbols x that appear exactly k times in $\bar{x}$. Finally, a total probability of

$$\frac{\varphi_1}{n} \tag{2.16}$$

is assigned to all the symbols in Ω that have never appeared in the sequence $\bar{x}$. Here ${\varphi'}_k$ is the smoothed values of the corresponding spectrum element φ_k. Smoothing is primarily necessary to prevent unobserved symbols being assigned a probability of zero. Several smoothing methods are available but will not be considered here for brevity of exposition. For the example of the African safari, the spectrum is readily calculated to be

$$\varphi_k = 1, \quad k = 1, 2, 3 \tag{2.17}$$

$$\varphi_k = 0, \quad k = 4, 5, 6. \tag{2.18}$$

So, if Good and Turing had been on the African safari, the probabilities assigned would have been

Zebra	0.25
Elephant	0.25
Giraffe	0.25
Other Animals	0.25.

(2.19)

Observe the significant divergence of the Good-Turing probability estimation (Equation (2.19)) with the empirical one (Equation (2.13)) and the Laplacian one (Equation (2.14)). The probability estimation of Good-Turing has been found to perform very well in practice in estimating the probability of "rare symbols," i.e., symbols that appear infrequently.

On the other hand, however, the Good-Turing estimator has resisted a satisfactory mathematical understanding so far[1] and is widely accepted to be "cryptic."

Recently, Orlitsky et al. have proposed a natural framework in which one can study probability estimation problems in this regime: small values of n compared to the size of the vocabulary $|\Omega|$. They show, using this framework, that the ML and Laplacian estimators perform quite poorly, and while the

[1]Turing died shortly after WWII and in the paper that Good published shortly after [24], he wrote that "Turing had demonstrated an intuitive explanation for the estimator, but I have forgotten it now".

Good-Turing estimator performs admirably well, it is not the best. Further, they demonstrate an "asymptotically optimal" probability estimator motivated by their framework. These results appear in a sequence of papers by Orlitsky et al. [45, 46, 47, 48, 49] and the rest of this section is a brief survey of their main results with respect to the context described so far.

To describe their framework, we take a brief detour to *information theory* in the next section, where we see the natural connection between probability estimation and lossless data compression. Much of the next section is my summary of Chapter 5 of [53].

2.1.7 Probability Estimation and Data Compression

Lossless data compression also starts with the same data sample as the probability estimation problem we have considered so far: Let

$$X_1, \ldots, X_n$$

be a collection of i.i.d. random variables drawn from an unknown distribution $\mathbb{F}$ over the vocabulary Ω. The data compression problem asks for an *injective* mapping of this observation into a *bit sequence*:

$$f_n : \Omega^n \mapsto \{0,1\}^* ,$$

such that the *length* of the bit sequence is as small as possible, with the constraint that the bit sequence can be used to exactly reproduce the original observation. In other words, the mapping f_n should be invertible. As engineering problems go, the data compression problem and the probability estimation problem are rather different. However, there is a very close connection between the two; this is a central result from information theory.

The central observation that connects these two problems is that good probability estimation is *sufficient* for good data compression. To see this, suppose that we have first learned the unknown distribution $\mathbb{F}$ (via good probability estimation). Then, a procedure known as the *Huffman* code lets us find the encoding f_n, that is optimal in the sense it minimizes the *expected* length of the resulting binary representation. Essentially, the length of the binary representation of a sequence $\bar{x} = (X_1, \ldots, X_n)$ is *inversely proportional*

to how likely the sequence is: the more likely sequences are represented by shorter bit sequences. Specifically, the length of the binary representation corresponding to a sequence $\bar{x}$ using the Huffman code is

$$\lceil -\log_2 \mathbb{P}[\bar{x}] \rceil, \tag{2.20}$$

where $\mathbb{P}[\bar{x}]$ is the probability of the sequence $\bar{x}$ with respect to the underlying distribution $\mathbb{F}$. Obviously, this entails knowing the distribution $\mathbb{F}$. The average representation length, averaged over the underlying distribution, is

$$nH(\mathbb{F}) = -\sum_{\bar{x} \in \mathcal{X}^n} \mathbb{P}[\bar{x}] \log_2 \mathbb{P}[\bar{x}], \tag{2.21}$$

where $H(\mathbb{F})$ is called the *entropy* of the underlying distribution.

Thus we see that good probability estimation allows us efficient data compression. The converse is true as well: if we have a good data compression scheme, then we can assign a probability of

$$2^{-l(\bar{x})} \tag{2.22}$$

to a sequence $\bar{x}$ which has representation length $l(x)$ under the data compression scheme. This probability estimator is now guaranteed to closely approximate the underlying distribution, if the data compression algorithm is close to being optimal. We conclude that the data compression and probability estimation, while different engineering problems, are really closely tied to each other.

2.1.8 Universal Data Compression

The scenario common in data compression is that the underlying distribution $\mathbb{F}$ is *unknown*; then we are in the realm of universal data compression. A common measure of how good a universal data compression algorithm performs is *redundancy*: this is the extra number of bits needed to represent the source when compared to the situation where the underlying distribution $\mathbb{F}$ is already known. The main result in information theory is that universal data compression algorithms exist that have *diminishing redundancy*, i.e., the number of extra bits needed grows *sublinearly* with the sample size; in

fact it only grows *logarithmically* with the sample size n: the extra number of bits is approximately (see Equation (2) of [45])

$$\frac{|\Omega| - 1}{2} \log\left(\frac{n}{2\pi}\right). \tag{2.23}$$

Practical data compression algorithms such as `compress` and `gzip` are provably asymptotically optimal data compression methods. As can be seen from Equation 2.23, universal data compression is efficient when the sample size (n) is large and the vocabulary size ($|\Omega|$) is small. This is usually the case with compressing binary files and hence the efficiency of `gzip` and `compress`.

On the other hand, when n is small and/or $|\Omega|$ is large, universal data compression is infeasible. This is the scenario under which the Laplacian and Good-Turing estimators fit in. To study this scenario, Orlitsky et al. [45] suggest studying some simpler properties of the observation: *patterns* and *profiles*; their main result is on universal compression of patterns and profiles of the observed sequences. Due to the close connection between data compression and probability estimation, their results shed insight into the context of the discussion in the previous section. We summarize their main results and the corresponding insights in the next section.

2.1.9 Patterns, Profiles and Good-Turing

Consider the observed sequence $\bar{x} \in \Omega^n$. Let the rank of a symbol x that appears in the sequence x be one more than the number of distinct symbols that have preceded x. The pattern of $\bar{x}$ is simply the concatenation of the ranks of the components $x_1, \ldots, x_n$ of $\bar{x}$. As an example, suppose Ω is the English alphabet with 26 letters and the observed sequence is

a b r a c a d a b r a.

The corresponding pattern is

1 2 3 1 4 1 5 1 2 3 1.

The profile is

2 2 0 0 1 0 0 0 0 0 0, (2.24)

i.e., there are two singletons ("c" and "d"), two doubletons ("b" and "r") and one symbol that appears five times ("a"). The main result of Orlitsky et al. is that while the individual sequences coming from a very large vocabulary may not be universally compressible, their *patterns* and *profiles* are: in particular, the redundancy $\mathcal{R}$ of compression of the pattern (and profile) of a sequence is (Section 2 of [45]):

$$(1.5 \log_2 e)\, n^{1/3}(1 + o(1)) \leq \mathcal{R} \leq \left(\pi\sqrt{\frac{2}{3}} \log_2 e\right) \sqrt{n}. \tag{2.25}$$

Here, $o(1)$ denotes a function that is going to zero as n grows large. So, the redundancy is only growing sub-linearly with the sample size n and thus patterns and profiles are universally compressible.

Motivated by these results on data compression, we can consider the probability estimation not of the sequence itself, but its *pattern*. Observe that both the Laplacian and Good-Turing estimators can also be viewed as pattern estimators. Borrowing the performance measure of redundancy from the data compression problem, Orlitsky et al. measure how good the Laplacian and Good-Turing estimators are. Specifically, they show the following results: For the Laplacian estimator, they show (Theorem 1 of [49])

$$\mathcal{R}_{\text{Laplace}} = \infty. \tag{2.26}$$

The following example demonstrates why the Laplacian estimator performs so C: consider the pattern $123 \ldots n$. This pattern represents a sequence where *every* element is "new". The Laplacian estimator assigns a probability of

$$\frac{1}{1} \cdot \frac{1}{3} \cdots \frac{1}{2n+1} = \frac{2^n \cdot n!}{(2n+1)!}, \tag{2.27}$$

which goes to zero as n grows. On the other hand, if the observation consisted of the DNA sequences of animals, then it is expected that every symbol is "new." With this underlying distribution, the correct probability assignment should be 1. Thus, the Laplacian estimator can perform arbitrarily poorly.

In contrast, the Good-Turing estimator has *bounded* redundancy (Theorem 3 of [49]): the redundancy per symbol $\mathcal{R}_{\text{GT}}$ is upper and lower bounded as

$$\log_2(1.39) \leq \mathcal{R}_{\text{GT}} \leq 1. \tag{2.28}$$

So, while the Good-Turing estimator cannot perform arbitrarily worse, it still has a finite redundancy. Motivated by their results on universal compression of patterns, Orlitsky et al. demonstrate that a slightly modified version of the Good-Turing estimator has zero asymptotic redundancy. The modification involves a slight change to the probability estimation in Equation 2.15.

$$\frac{c \max(\varphi_{k+1}, n^{1/3})}{\varphi_k}, \quad k = 1 \ldots n. \tag{2.29}$$

Similarly, Equation 2.16 is modified as

$$c \max(\varphi_1, n^{1/3}). \tag{2.30}$$

With this modification, Theorem 4 of [49] claims that the per symbol redundancy is sub-linear:

$$\mathcal{R}_{\text{modifiedGT}} \leq O(n^{2/3}). \tag{2.31}$$

2.1.10 Discussion

The main contributions of Orlitsky et al. could be viewed as providing a common framework (information-theoretic) to study very old probability estimation problems. This framework has allowed them to quantitatively characterize the performance of Laplacian and Good-Turing estimators. It has also allowed them to come up with simple modifications that allow strong improvements over the Good-Turing estimator. Overall, this is very much an active area of research both theoretically and from a practical viewpoint. In particular, the key theoretical question of the best probability estimators of the pattern, in the sense of smallest redundancy, is still open. The design of corresponding estimators is also open. On the practical side, applications to computational linguistics and data mining of these new ideas are as yet completely unexplored. In a recent work, Orlitsky et al. [54] have applied their improved estimation methods to a benchmark problem in data mining: text classification. They show that binary text classification based on the modified Good-Turing estimator (cf. Equation (2.29)) is an improvement over that based on the Laplacian (and `add-constant` estimators in general).

While this improvement is not that surprising, a more relevant comparison would be between classification methods based on regular Good-Turing

and the modified one. The following rule of thumb analysis demonstrates that perhaps the modified Good-Turing method would have little to offer (or, perhaps even worse) in typical natural language data settings. Typical documents have total number of words on the order of about 1000. (This seems to be true even if the stop words are removed, although this is not explicitly mentioned in the analysis in [54].) This means that

$$n^{\frac{1}{3}} \approx 10. \tag{2.32}$$

On the other hand, the frequency of each rare word is typically no more than 5 or so (Good-Turing smoothing is typically considered only for such low frequency words). So, the max operation in Equation (2.29) for most words in the document will result in 10, a constant that is independent of the word! This back-of-the-envelope analysis suggests that more work is needed to make the modified Good-Turing estimator a practical success.

2.1.11 Unseen Element Size Estimation via Probability Estimation of Rare Events

The approach of Orlitsky et al. (described in Section 2.1.6) suggests a natural approach to the problem of interest in this chapter: estimation of the number of unseen elements. Fix the data set of size n and denote the profile by $\varphi_1, \ldots, \varphi_n$. We can ask for the probability distribution that yields the largest likelihood of the profile vector $\varphi_1, \ldots, \varphi_n$. This is the so-called *maximum likelihood* (ML) probability distribution and "best explains" the observed profile. The cardinality of this ML probability distribution is then taken as the size of the underlying vocabulary. Since the number of seen elements is directly known, the number of unseen elements is then readily found.

This is a very well posed estimator of the number of unseen elements. But it suffers from two fundamental drawbacks:

- Practically, the computation of the ML probability distribution appears to be extremely hard. Indeed, even for n of only 5 or 6, this appears to be nearly impossible [48]. Unless this issue is addressed, there is little scope of applying this method in practice (where n is in the tens of hundreds if not more).

- There is no theoretical performance guarantee available for using the ML probability distribution to evaluate the unseen element size. Such a result would have set a baseline performance guarantee that would suggest the potential suitability of devoting effort to further study this approach.

2.2 Automatic Fluency Assessment

Previous work in this area can be studied with regard to two aspects of quantifying fluency:

1. the choice of quantifiers of fluency, and
2. the methods of measuring the quantifiers and, hence, the automated methods of language proficiency assessment.

Although perceptions of language fluency are largely subjective, several studies have aimed at exploring human-rated assessment of language fluency towards developing the right quantifiers. These studies sought to quantify perceived fluency in terms of objective properties of speech by examining the correlation of a set of quantifiers of speech production with human-assigned fluency scores. In particular, the explorations have involved examining the relation between temporal and lexical features of speech, and fluency scores.

One of the early studies in quantifying fluency is by Lennon in [55], which was done in the context of a longitudinal study of finding factors that affected oral fluency when the subjects' exposure to the second language increased. Comparing speech at the beginning and the end of a six-month stay of four speakers in an English-speaking country, he found that these subjects were perceived to be more fluent. He attributed the improved fluency scores to increase in speaking rates and decrease in filled-pauses. In a similar context, the study by Riggenbach [56] examined the relationship between a set of temporal measures of speech and perceptions of fluency while assessing second-language English speech of six Chinese speakers. This study found that the classification of speakers as being fluent or non-fluent by their instructors was strongly influenced by the number of unfilled pauses. A later study by Kinkade [57] in a similar setting sought to include syntactic complexity of the utterances in addition to the temporal features of speech rate

and mean length of unfilled pauses. More recently, expanding the set of quantifiers to include accuracy features via morpho-syntactic error rates, Mizera [58] observed that the set of quantifiers used in a multiple regression model for approximating the fluency scores accounted for nearly 83% of the variance in fluency scores.

Several studies have sought to assess the effect of vocabulary knowledge (via lexical richness) on language proficiency. The term lexical richness is intended to cover the following different aspects of lexical use [59]:

- lexical diversity, which is the variety of active vocabulary deployed by a speaker or writer [60],
- lexical sophistication, the number of low frequency words,
- lexical density, which is the proportion of content words in the total words used [61].

The measures of lexical richness available in the literature can be broadly classified as follows:

- Word-list free measures. This set of measures is obtained without a dictionary-based list of words. Since the frequency information of words is not taken into consideration, these measures are regarded as focusing on lexical diversity. The important measures are:

    1. The type-token ratio (TTR), given by the ratio of number of word types to the number of word tokens: owing to the effect of the number of tokens on the measure, this measure is widely considered inadequate for quantifying lexical richness.
    2. The Guiraud index, given by the ratio of the word types to the square-root of the word tokens: this is used as an alternative to the TTR, since the dependence on the number of tokens is sublinear as opposed to being linear. Nevertheless, it can still be affected by widely differing counts in the number of tokens.
    3. The d(iversity) measure D [60], a parameter that captures the deviation of the TTR-curve of the sample from that of a theoretical TTR model.

- Word-list based measures. These measures first distinguish words based on their frequency of occurrence in the language before applying word-list free measures.

    1. The advanced Guiraud index, a variant of the Guiraud index for advanced word types.
    2. A derived form of the limiting relative diversity (LRD) given by $\sqrt{D(verbs)/D(nouns)}$.
    3. Lexical frequency profile (LFP) gives the percentage of words a learner uses at different vocabulary frequency levels. By *frequency level* is meant a class of words (or appropriately chosen word units) that are grouped based on their frequencies of actual usage in corpora. P-Lex [62] is another approach that uses the frequency profile of the words to assess lexical richness.

Analyzing essays written by second language learners of English, Laufer and Nation in [61] have shown that LFP correlates well with an independent measure of vocabulary knowledge and that it is possible to categorize learners according to different proficiency levels using this measure.

Quantitative measures of vocabulary richness in semi-spontaneous speech and their correlations with scores of language proficiency have been studied in [63]. The results of this study showed high correlations (greater than 0.70) between the measures LRD, Guiraud Index, advanced Guiraud's index and the D-measure with scores of language proficiency. This suggests the suitability of these measures for capturing lexical richness. Another study by Daller and Xue [59] investigated the use of a set of word-list free measures (TTR, Guiraud index and D) and a set of word-list based measures (LFP, Advanced Guiraud's index and P-Lex) for distinguishing oral proficiency levels of second language learners. They found that the Guiraud index and D-measure best captured the difference in proficiency levels.

In an effort to investigate the effects of various aspects of speech on perceptions of fluency, Kormos and Dénes sought to find the extent to which effort of production (quantified by the temporal measures such as speech rate), command of grammar (measured in terms of the ratio of number error-free clauses to the total number of clauses) and use of vocabulary (by measuring lexical richness using the D-measure and the number of word tokens) affected

perceptions of fluency. They concluded that perceptions of fluency were more strongly correlated with the temporal variables of speech than with measures of lexical richness and grammatical accuracy [64].

An important contribution of these studies has been identifying a set of quantifiers representing the temporal measures of perceptions of fluency as well as those representing lexical richness as indicators of language proficiency. No less a contribution has been the understanding that the subjective notion of fluency can be quantified to a greater extent by temporal measures of speech and to a lesser extent by lexical measures and measures of grammatical accuracy. A salient feature of the studies mentioned above is that the quantifiers chosen to represent perceptions of fluency and language proficiency were obtained *manually* from the utterances as well as their transcriptions.

As an essential step toward automatic assessment of fluency, it was then imperative to explore the possibility of obtaining the quantifiers of fluency *automatically* from the speech segment. In addition to obtaining the measures automatically, the extent to which measurements thus obtained correlate with human-rated language fluency scores would have to be considered. Research along these lines has shown that a set of temporal measures of speech obtained automatically are good quantifiers of perceived fluency in both spontaneous and read speech [65, 66]. In the case of spontaneous speech in Dutch as a second language, they showed that a set of automatic measures including rate of speech, phonation-time-ratio and length of pauses are good indicators of perceived fluency. The other quantifiers explored in this study were mean length of runs, frequency of filled-pauses and articulation rate. Automatic measurement of the quantifiers in their study was done by the use of an automatic speech recognizer(ASR) trained on non-native speech.

More recently, Yoon [67] studied an automatic scoring model for fluency in second language speech based on automatically extracted features. This study used a set of temporal and syntactic features (shown in previous studies [55, 56, 57, 65, 58] to have significant correlation with L2 fluency scores), automatically obtained by the use of ASR, in a multiple regression model for approximating human assigned fluency scores. The study concludes that features of syntactic complexity showed the least correlation with fluency scores, whereas quantifiers of speed (specifically, rate of speech and mean length of runs) showed the highest correlation with fluency scores.

Table 2.1: Quantifiers used in the preliminary scoring model of the *Speechrater* to assess fluency, lexical use, pronunciation and grammatical accuracy. Mean deviation is the mean of the absolute differences between the quantifier and its mean value.

Criterion	**Description**
Fluency	Average chunk length in words articulation rate (in words per second) mean deviation of chunks in words total duration of silences/no. of words mean silence duration mean duration of long pauses frequency of long pauses/no. of words
Vocabulary richness	No. of unique words per second No. of unique words/total duration of utterance
Pronunciation	Global HMM acoustic model score (normalized)
Grammar	Global language model score (normalized)

We have identified some of the quantifiers of fluency that have been studied and some attempts to measure them automatically using ASRs. The next step towards automatic language proficiency assessment is the design of a complete system. The state-of-the-art system for assessing language proficiency is the *SpeechRater*, an automated scoring system for spontaneous speech of Engish learners used operationally in the TOEFL Practice Online assessment [19]. The system is intended to provide a platform for scoring spontaneous non-native speech for comprehensibility, coherence and appropriateness. A prototype of the system is currently being used by the Educational Testing Service (ETS) for low-stakes assessment in the form of preparation for online test takers. It has a feature extractor that uses the output of a speech recognizer to generate quantifiers of fluency, lexical use, grammar and pronunciation. A select set of these quantifiers are then combined in a multiple regression model to generate the final proficiency score. The quantifiers chosen and the representative criteria are outlined in Table 2.1. The study [19] reports experiments performed using field study data comprising of two data sets.

As can be seen from the table, the features chosen are mostly quantifiers of fluency and only marginally represent aspects of lexical use, pronunciation and grammatical accuracy. However, the set of quantifiers was further refined to obtain the final set of quantifiers that are used in the scoring model. This

set comprises the quantifiers *global HMM acoustic model score*, *articulation rate*, *number of unique words per second*, *average chunk length in words* and *global language model score* with weights 4, 2, 2, 1 and 1 respectively, chosen to cover different aspects of speaking ability with their relative importance in the scoring procedure as perceived by domain experts. The multiple regression correlations between human-assigned and machine-assigned scores for different data sets are 0.57 and 0.68 and their corresponding κ scores indicating human-computer agreement are 0.51 and 0.61. Zechner et al. remark that the performance is inadequate for high-stakes testing purposes and that this system is still under experimentation.

2.2.1 Thin-Slices Assessment

A series of studies in experimental social psychology by Ambady et al. [68] have sought to investigate the rapid, unwitting and impressionistic judgments that people make about certain behavioral characteristics of others. In particular, these studies are concerned about the extent to which people's impressions and behavior are influenced by such rapid judgments, the accuracy of judgments made so quickly and the bases upon which such judgments are made. A brief excerpt of the expressive behavior that is sampled from the behavioral stream has been termed *a thin-slice*. They in turn, cite studies that show empirical evidence supporting the idea that "a brief acquaintance often does result in amazingly rich impressions based on cues that are derived entirely from expressive movements— from appearance, gesture and manner of speaking."

Summarizing results from their own as well as those of other studies, Ambady et al. discuss the extent to which thin-slice judgments can be predictive about outcomes in diverse areas of social life ranging from performance in educational, organizational and health-care settings, to aspects of interpersonal relationships and individual differences such as sexual orientation. The thin-slices typically are of the order of few seconds and judges with different cultural backgrounds and degrees of association with the target person were considered.

One of the observations of the study is that the reliability of the thin-slice judgment made was highly dependent on the nature of the criterion being

judged. Accordingly they note that the observability of the criterion being judged (such as impressions about the person being active or competent) influences reliability of the thin-slice judgement. Moreover, they also observe that judgment reliability is affected by the manner in which the behavioral stream is observed. For instance, their results suggest that judgments such as whether a candidate appears confident, likable or active, are more consistent when judged via video clips than via audio clips. They also report that for the attributes considered, reliability is greater on average for judgments of clips that include both audio and video than for those based on silent video clips; silent video clips were judged more reliably on average than audio clips, which in turn were judged more reliably than content-filtered speech.

2.2.2 Summary of Previous Work in Fluency Assessment

We saw in this section that temporal aspects of speech have been shown to correlate well with fluency scores of spontaneous speech. In addition we saw that aspects of grammatical accuracy and lexical use are not as good predictors of language proficiency as those of temporal aspects of speech. As such, we notice that the correlations of different measures of lexical richness with fluency have not yet been studied.

Another notable observation regarding previous studies on automatic assessment is that the objective measurements mentioned above were obtained by using automatic speech recognizers (ASRs) trained using language specific data. While building an ASR may seem like a first approach to automate aspects of language testing using features obtained from the speech signal, the design of one such system is resource-intensive. Training an ASR calls for a large amount (typically several tens of hours) of high-quality speech data recorded under noise-free conditions. It also involves having the corresponding transcriptions of the speech segments available. The speech recognition system in [66], for instance, used 200 hours of native speech for training and that in [19] about 150 hours of second language speech. Collecting such a corpus may be infeasible for various reasons. For instance, if the language spoken is a minority language such resources may be hard to find. Moreover, the performance of an ASR is strongly influenced by the quality of the recordings. In the event that the data was recorded in noisy environments,

it may not be suitable for use with an ASR. In addition, second language speech with its inherent lexical and syntactic errors and wide accent variations makes speech recognition challenging. Any or all of these can result in imperfect accuracy levels in current ASR systems, rendering them as yet inadequate for automatic assessment. This motivates the search for alternative methods of automatic measurement of quantifiers of language fluency. One of the goals of this study is to find alternative methods of fluency assessment with the ability to deal with most (if not all) of the shortcomings outlined above.

Studies based on thin-slice judgments have considered several facets related to social life. Language competence (in particular, oral fluency in a second language) is not far removed from this domain. While the cognitive aspects of human assessment of oral fluency (and thin-slice judgments of fluency) are beyond the scope of this study, the objective measures of perceptions of fluency are central to this study. With the intent to find out if snippets of an utterance bear the information necessary for judgments of oral fluency, we perform a thin-slice assessment of fluency using quantifiers that are obtained from a random snippet of the spontaneous utterance.

With the material presented thus far as the background, we are now ready to formalize the goal of this study with regard to automatic fluency assessment. This study is an attempt to address the following research goals:

1. to understand the extent to which a set of quantifiers of lexical use correlate with human-assigned fluency scores;

2. to make signal level measurements leading to a set of quantifiers of temporal aspects of speech production and verify that they quantify human perceptions of fluency reasonably well;

3. to design an end-to-end automatic fluency assessment system using a good set of quantifiers of fluency and evaluate its performance on an available data set; and,

4. to study the degree to which automatic assessment of oral fluency based on a random short snippet of the entire utterance agrees with that based on the entire utterance.

The experiments that we performed toward addressing the research questions above are discussed in Chapter 4 of this thesis.

2.3 Prediction of Click-Through Rates of Advertisements

The problem of modeling advertisement click-through rates by generating a set of features and then using them in a logistic regression model has been studied by [69]. Several approaches to variable selection have been proposed which can be broadly classified as those that perform variable ranking and those that perform forward or backward selection of nested subsets [22]. In the current work, we take the forward selection approach to perform the task of variable selection sequentially. Accordingly, we first perform variable selection and then use the variables thus obtained to fit a logistic regression model to predict the probability of a click on an advertisement. The details of our experiments are discussed in Chapter 5.

CHAPTER 3

KNOWING THE UNSEEN: ESTIMATING VOCABULARY SIZE OVER UNSEEN SAMPLES

3.1 Introduction

A phenomenon characteristic of natural language corpora is that of *skewed* frequency distributions: a large number words occur with very low frequencies (and a small number of words occur with very high frequencies). Consider Figure 3.1, which plots the number of words having different frequencies of occurrence from the British National Corpus, involving nearly 100 million words. From the figure, the phenomenon we have mentioned is very clear: a vast majority of the words occur with very low frequencies (in the single digits). To drive this point home, consider a very different kind of corpus (in many ways): the works of Shakespeare. The numbers of words occurring fewer than 10 times are tabulated in Table 3.1: we see that nearly all the words Shakespeare ever used occur less than 10 times. Furthermore, nearly half of all the words used by Shakespeare occur just once.

Let us now digress a bit and consider the distribution of heights of adults in a population. We know that heights are normally distributed, meaning that we find very few extremely tall and extremely short people compared to a sea of people with normal heights. Contrasted with this, a population with distributions similar to word frequency distributions have a coterie of giants, some shorter people and an army of dwarfs.

This phenomenon of skewed frequency distributions is not necessarily specific to words. Syllables (components of words) and their frequencies also display the same phenomenon. In particular, this is common in languages with complex syllable structure such as German and English. In these languages, a few hundred syllables account for most syllables uttered. On the other hand, a majority of syllables are rarely used. We have moved from words to syllables, and moving further down to duration in speech sounds

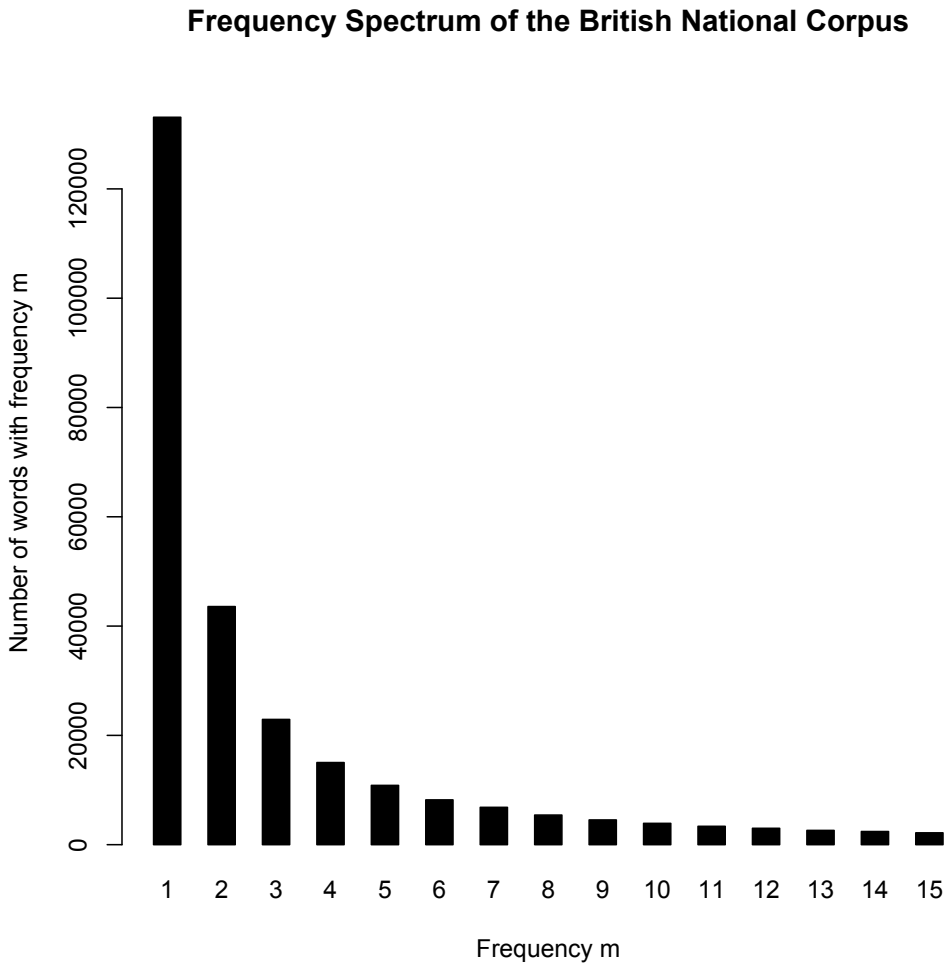

Figure 3.1: Plot of number of words as a function of the number times they occur in the BNC standardized corpus.

Table 3.1: Number of words Shakespeare used as a function of the number of times they occurred. Nearly half of all the words used by Shakespeare appear just once.

Frequency	Number of words
All	31,534
1	14,376
2	4343
3	2292
4	1463
5	1043
6	837
7	638
8	519
9	430
10	364

(which are components of syllables), it has again been observed that most of the features quantifying aspects of duration occur only a few times. On the other hand, a small set of duration features are the ones that are most commonly observed. A good reference to see these observations in detail is available in [70]. This same phenomenon also occurs in a variety of settings that are quite removed from natural languages. For instance, a vast number of queries to large databases (such as the one Microsoft's `SQL server` handles) occur just a few times [71].

The law of large numbers guarantees that in a population with finite vocabulary, the ML estimates of probabilities of occurrence of the individual words converge to population probabilities when we consider large enough samples. As a consequence of this, at a large enough sample size we see that all the words in the vocabulary have been sampled at least once and continuing the sampling procedure only increases the frequency of an individual word with no more "new" words.

When vocabulary is not finite, however (as in all natural language corpora), ML estimates of population probabilities are misleading; this is compounded by the fact that many words (those belonging to the open class of words or those occurring as neologisms, to name a few) have a small chance of occurring in any given sample. This results in increasingly larger samples with increasing vocabulary sizes even for very large sample sizes. Such sample sizes where the vocabulary size is still increasing and where the low-frequency

components of the spectrum are non-negligible are said to be located in the *large number of rare events* (LNRE) zone [37]. Given that a huge number of elements occur just a few times, it is highly likely that a good number of elements have never been seen. So, the problem of estimating the number of unseen elements is an interesting and challenging one in the LNRE zone, the region of interest for our study.

In the natural language context, we saw in Section 1.1 that it may be of particular interest to estimate the total most likely vocabulary size of the population from which the corpus was drawn. Since the total vocabulary size is just the sum of the seen and unseen vocabulary sizes, this is the same as the problem statement from earlier. The total vocabulary size could be useful for:

- comparing corpora, creating language models and making generalizations about specific linguistic phenomena in a language;
- choosing an optimal vocabulary size which will then influence the building of training data for designing speech recognizers.

Answering this question is the entire focus of this chapter. We propose a nonparametric estimator of the number of unseen elements in a sample that is characterized by a large number of events with small chance of being seen. Our estimator is nonparametric and novel. It distinguishes itself from the existing literature in the following important way:

> We can prove the consistency of the estimator in the context of a natural probabilistic model.

Further, from the theoretical angle, we also analyze its rate of convergence. On the practical side, we see a favorable comparison of our estimator with the state-of-the-art (both parametric and nonparametric) ones on several standard natural language corpora. Our main experiment involves computing the performance of the different estimators (including ours) in extrapolating vocabulary size. We show that in extrapolating vocabulary sizes over sample sizes about *twice* the observed sample, performances comparable to those of the state of the art are achieved.

This chapter is organized as follows. We first consider a nonparametric estimator constructed using the widely used Good-Turing estimator of probability of unseen elements in the state of the art: the so-called *Good-Turing*

estimator of vocabulary size (Section 3.2). We see a consistency property of this estimator in a natural probabilistic model of LNRE. While this consistency property is straightforward, it appears to be new to the literature. This consistency property of the Good-Turing estimator is in a very narrow LNRE context: the situation when all rare events take on *exactly* the same small probability of occurrence. Allowing variability among the probability of words, rare they all be, we see that the Good-Turing estimator is no longer consistent. This is completed in Section 3.2 which sets the stage for our estimator, the topic of Section 3.3. Our main result is the consistency of our novel estimator in the general LNRE context. In this sense, it can be viewed as a natural generalization of the Good-Turing estimator which was consistent only with the special case of uniform probabilities in the LNRE context.

The secondary aspects of the consistency of the estimator are explored in Section 3.4: uniform convergence and rates of convergence are discussed here. Section 3.5 covers the experiments on several standard large natural language corpora.

3.2 Good-Turing Estimator

3.2.1 The Estimator

Following the terminology in [27], the Good-Turing estimator of the total vocabulary size is

$$V_{\mathrm{GT}} = \frac{V}{1 - \frac{\varphi_1}{n}} = \frac{\sum_{k=1}^{n} \varphi_k}{1 - \frac{\varphi_1}{n}}. \tag{3.1}$$

It is simply stated and is readily derived from the Good-Turing formula for the total probability of the number of unseen words: $\frac{\varphi_1}{n}$. This particular formula for the total probability of the number of unseen words first appeared in an article by I. J. Good [52] where he originally attributed it to A. M. Turing. In [52], Good showed intuitively that this estimator of probability mass is unbiased (this fact was shown rigorously by Robbins in [72]). The interpretation of Equation (3.1) from this probability expression is straightforward. The denominator of Equation (3.1) represents the total probability of the words that are seen and, indeed, this is approximately the right quan-

tity by which the seen vocabulary size is a fraction of the total vocabulary size (hence termed as a coverage based estimator [26]). Bunge and Fitzpatrick [26] mention the applicability of the estimator of probability mass of unseen words even in applications where the underlying distribution is not known although it was proposed under the assumption of equiprobable classes. Gandolfi and Sastri indicate that they "stretch the meaning" of the estimate of the coverage of a sample to arrive at the Good-Turing estimator of vocabulary size.

The Good-Turing estimator for the total vocabulary size, while widely used (refer to Section 2.1.2 where we discuss this in the context of the set of estimators), has (remarkably) no theoretical properties proved. (Several theoretical properties are known in terms of the Good-Turing estimator when used for the total probability of the number of unseen words [72, 73, 74], however.) It has been found empirically that the estimator performs poorly when the primary assumption of equiprobable classes is violated [27]. We provide a simple theoretical property of the Good-Turing estimator that appears novel in the literature and corroborates the empirical finding of underestimation from available literature and from our own experiments.

3.2.2 Uniform Rare Events and Consistency

Consider the probability model where the probabilities of the words are the same (uniform probability measure), but the actual quantity changes with the sample size. This means that we have a sequence of vocabularies Ω_n and accordingly a sequence of probability measures $\mathbb{P}_n$, all indexed by the sample size n. The probability of any word is the same, say

$$\mathbb{P}\left[\omega \in \Omega_n\right] = \frac{c}{n}, \quad \forall \omega \in \Omega_n. \tag{3.2}$$

This simple model is said to be an example of the so-called LNRE probability models. This means that the vocabulary size is

$$|\Omega_n| = \frac{n}{c}. \tag{3.3}$$

We claim the simple consistency result of the Good-Turing estimator of the total vocabulary size (cf. Equation (3.1)).

Proposition 1

$$\lim_{n\to\infty} \frac{||\Omega_n| - V_{\mathrm{GT}}|}{n} = 0, \quad \textit{almost surely.} \tag{3.4}$$

Proof: We begin with the chance that any word occurs just once. It is

$$n \cdot \frac{c}{n}\left(1 - \frac{c}{n}\right)^{n-1}. \tag{3.5}$$

This is because the single occurrence can occur in any of the n positions and the second and third terms represent the chance of occurring once and not occurring the remaining times, respectively. So, the expected number of singleton words is

$$\mathbb{E}\left[\varphi_1\right] = \frac{n}{c} \cdot c\left(1 - \frac{c}{n}\right)^{n-1}. \tag{3.6}$$

Now making use of the fact that $\lim_{n\to\infty}\left(1 - \frac{c}{n}\right)^{n-1} = e^{-c}$, we see that

$$\lim_{n\to\infty} \frac{\mathbb{E}\left[\varphi_1\right]}{n} = e^{-c}. \tag{3.7}$$

More generally, the same calculation yields for any fixed $k < n$

$$\lim_{n\to\infty} \frac{\mathbb{E}\left[\varphi_k\right]}{n} = \frac{c^{k-1}}{k!}e^{-c}. \tag{3.8}$$

We will see shortly, as an application of a more general result in Section 3.3.5, that

$$\lim_{n\to\infty} \frac{\varphi_k}{n} = \frac{c^{k-1}}{k!}e^{-c}, \quad \text{almost surely.} \tag{3.9}$$

The Good-Turing estimator (from Equation(3.1)) of the total vocabulary size (normalized by the sample size)

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n}\varphi_k}{\left(1 - \frac{\varphi_1}{n}\right)n} = \frac{\sum_{k=1}^{\infty}\frac{c^{k-1}}{k!}e^{-c}}{1 - e^{-c}} \tag{3.10}$$

$$= \frac{\frac{e^{-c}}{c}\left(e^{c} - 1\right)}{1 - e^{-c}} \tag{3.11}$$

$$= \frac{1}{c}. \tag{3.12}$$

Equation (3.10) used the step derived earlier in Equation (3.9). (There is still a slight technicality required to ensure that while the numerator and denominator separately converge, their ratio converges too, but this will follow from the more general result, Theorem 1, claimed in the next section.) Since the total actual vocabulary size normalized by the sample size also converges to $\frac{1}{c}$ (cf. Equation (3.3)), this completes the proof.

3.2.3 Nonuniform Rare Events and Inconsistency

Now consider the scenario when the underlying probability distribution is a *binary mixture* of two separate uniform LNRE distributions:

- the probability of any word is either $\frac{c_A}{n}$ or $\frac{c_B}{n}$;
- there are a total of $\frac{n}{2c_A}$ words of probability $\frac{c_A}{n}$;
- there are a total of $\frac{n}{2c_B}$ words of probability $\frac{c_B}{n}$.

So the mixture of the two uniform distributions is via a Bernoulli random variable with probability 0.5. Then Equation (3.3) generalizes to

$$\lim_{n\to\infty} \frac{|\Omega_n|}{n} = \frac{1}{2c_A} + \frac{1}{2c_B}. \tag{3.13}$$

Similarly, Equation (3.9) generalizes to

$$\lim_{n\to\infty} \frac{\varphi_k}{n} = \frac{1}{2}\frac{c_A^{k-1}}{k!}e^{-c_A} + \frac{1}{2}\frac{c_B^{k-1}}{k!}e^{-c_B}, \quad \text{almost surely.} \tag{3.14}$$

So the Good-Turing estimator (from Equation(3.1)) of the total vocabulary size (normalized by the sample size)

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n}\varphi_k}{\left(1-\frac{\varphi_1}{n}\right)} = \frac{\frac{1}{c_A}\left(1-e^{-c_A}\right)+\frac{1}{c_B}\left(1-e^{-c_B}\right)}{\left(1-e^{-c_A}\right)+\left(1-e^{-c_B}\right)}. \tag{3.15}$$

Comparing this to the actual total vocabulary size (cf. Equation (3.13)), we see that the consistency no longer holds (as long as $c_A \neq c_B$). In fact, we can see readily by very basic algebraic manipulations that the Good-Turing estimator of the vocabulary strictly *underestimates*, i.e.,

$$\frac{\frac{1}{c_A}\left(1-e^{-c_A}\right)+\frac{1}{c_B}\left(1-e^{-c_B}\right)}{\left(1-e^{-c_A}\right)+\left(1-e^{-c_B}\right)} < \frac{1}{2c_A}+\frac{1}{2c_B}, \tag{3.16}$$

whenever $c_A \neq c_B$.

Our main contribution starts with this observation and arrives at the appropriate expression that does indeed converge for all mixtures of uniform LNRE distributions. This is the focus of the next section.

3.3 Novel Estimator of Vocabulary Size

Our nonparametric estimator for the number of unseen elements is motivated by the characteristic property of word frequency distributions, the LNRE [37]. We also demonstrate that the estimator is strongly consistent under a natural scaling formulation described in [75].

3.3.1 A Scaling Formulation

Our main interest is in probability distributions $\mathbb{P}$ with the property that a large number of words in the vocabulary Ω are unlikely, i.e., the chance any word appears eventually in an arbitrarily long observation is strictly between 0 and 1. The authors in [37, 76, 74] propose a natural scaling formulation to study this problem; specifically, [37] has a tutorial-like summary of the theoretical work in [75, 76]. In particular, the authors consider a *sequence* of vocabulary sets and probability distributions, indexed by the observation size n. Specifically, the observation $(X_1, \ldots, X_n)$ is drawn i.i.d. from a vocabulary Ω_n according to probability $\mathbb{P}_n$. If the probability of a word, say $\omega \in \Omega_n$ is p, then the probability that this specific word ω does not occur in an observation of size n is

$$(1-p)^n .$$

For ω to be an unlikely word, we would like this probability for large n to remain strictly between 0 and 1. This implies that

$$\frac{\check{c}}{n} \leq p \leq \frac{\hat{c}}{n}, \tag{3.17}$$

for some strictly positive constants $0 < \check{c} < \hat{c} < \infty$. We will assume throughout this study that $\check{c}$ and $\hat{c}$ are the same for every word $\omega \in \Omega_n$. This implies

that the vocabulary size is growing *linearly* with the observation size:

$$\frac{n}{\hat{c}} \leq |\Omega_n| \leq \frac{n}{\check{c}}.$$

This model is called the *LNRE zone* and its applicability in natural language corpora is studied in detail in [37].

3.3.2 Shadows

Consider the observation string $(X_1, \ldots, X_n)$ and let us denote the quantity of interest — the number of word types in the vocabulary Ω_n that are not observed — by $\mathbb{O}_n$. This quantity is random since the observation string itself is. However, we note that the distribution of $\mathbb{O}_n$ is unaffected if one relabels the words in Ω_n. This motivates studying the probabilities assigned by $\mathbb{P}_n$ without reference to the labeling of the word; this is done in [76] via the *structural distribution function* and in [74] via the *shadow*. Here we focus on the latter description:

Definition 1 *Let X_n be a random variable on Ω_n with distribution $\mathbb{P}_n$. The* shadow *of $\mathbb{P}_n$ is defined to be the distribution of the random variable $\mathbb{P}_n(\{X_n\})$.*

As an example, suppose Ω_n is $\{a, b, c, d\}$ and

$$\begin{aligned} \mathbb{P}_n\{a\} &= \mathbb{P}_n\{b\} && (3.18)\\ &= \frac{1}{2}\mathbb{P}_n\{c\} && (3.19)\\ &= \frac{1}{2}\mathbb{P}_n\{d\} && (3.20)\\ &= \frac{1}{6}. && (3.21) \end{aligned}$$

Then, the shadow of $\mathbb{P}_n$ is a random variable that takes values $\frac{1}{6}$ and $\frac{1}{3}$ with probabilities $\frac{1}{3}$ and $\frac{2}{3}$, respectively. For the finite alphabet situation we are considering, specifying the shadow is *exactly equivalent* to specifying the unordered components of $\mathbb{P}_n$, viewed as a probability vector.

For the finite vocabulary situation we are considering, specifying the shadow is *exactly equivalent* to specifying the unordered components of $\mathbb{P}_n$, viewed as a probability vector.

3.3.3 Scaled Shadows Converge

We will follow [74] and suppose that the scaled shadows, the distribution of $n \cdot \mathbb{P}_n(X_n)$, denoted by Q_n converge to a distribution Q. As an example, if $\mathbb{P}_n$ is a uniform distribution over a vocabulary of size cn, then $n \cdot \mathbb{P}_n(X_n)$ equals $\frac{1}{c}$ almost surely for each n (and hence it converges in distribution). From this convergence assumption we can, further, infer the following:

1. Since the probability of each word ω is lower and upper bounded as in Equation (3.17), we know that the distribution Q_n is non-zero only in the range $[\check{c}, \hat{c}]$.

2. The "essential" size of the vocabulary, i.e., the number of words of Ω_n on which $\mathbb{P}_n$ puts non-zero probability can be evaluated directly from the scaled shadow, scaled by $\frac{1}{n}$ as

$$\int_{\check{c}}^{\hat{c}} \frac{1}{y}\, dQ_n(y). \tag{3.22}$$

   Using the dominated convergence theorem, we can conclude that the convergence of the scaled shadows guarantees that the size of the vocabulary, scaled by $1/n$, converges as well:

$$\frac{|\Omega_n|}{n} \to \int_{\check{c}}^{\hat{c}} \frac{1}{y}\, dQ(y). \tag{3.23}$$

3.3.4 Profiles and Their Limits

Our goal in this study is to estimate the size of the underlying vocabulary, i.e., the expression in (3.22),

$$\int_{\check{c}}^{\hat{c}} \frac{n}{y}\, dQ_n(y), \tag{3.24}$$

from the observations $(X_1, \ldots, X_n)$. We observe that since the scaled shadow Q_n does not depend on the labeling of the words in Ω_n, a *sufficient statistic* to estimate (3.24) from the observation $(X_1, \ldots, X_n)$ is the *profile* of the observation: $(\varphi_1^n, \ldots, \varphi_n^n)$, defined as follows. φ_k^n is the number of word types that appear exactly k times in the observation, for $k = 1, \ldots, n$. Observe

that

$$\sum_{k=1}^{n} k\varphi_k^n = n,$$

and that

$$V_n \stackrel{\text{def}}{=} \sum_{k=1}^{n} \varphi_k^n \tag{3.25}$$

is the number of *observed* words. Thus, the object of our interest is

$$\mathbb{O}_n = |\Omega_n| - V_n. \tag{3.26}$$

3.3.5 Convergence of Scaled Profiles

One of the main results of [74] is that the scaled profiles converge to a deterministic probability vector under the scaling model introduced in Section 3.3.3. Specifically, we have from Proposition 1 of [74]:

$$\sum_{k=1}^{n} \left| \frac{k\varphi_k}{n} - \lambda_{k-1} \right| \longrightarrow 0, \quad \text{almost surely,} \tag{3.27}$$

where

$$\lambda_k := \int_{\check{c}}^{\hat{c}} \frac{y^k \exp(-y)}{k!}\, dQ(y) \quad k = 0, 1, 2, \ldots. \tag{3.28}$$

This convergence result serves as a convenient tool to evaluate the performance of the Good-Turing estimator for the LNRE regime; we will see next that the Good-Turing estimator always strictly underestimates. More importantly, this convergence result also suggests a natural estimator for $\mathbb{O}_n$ (cf. Equation (3.26)).

3.3.6 Good-Turing Vocabulary Estimator Underestimates

Based on the convergence result in Equation (3.27) we see that the Good-Turing vocabulary estimate (scaled by sample size) converges almost surely in the LNRE regime as follows:

$$\frac{V_{\text{GT}}}{n} = \frac{\frac{1}{n} \cdot V_n}{1 - \frac{\varphi_1}{n}} \to \frac{\int_{\check{c}}^{\hat{c}} \frac{\left(1 - e^{-y}\right)}{y} \quad dQ(y)}{\int_{\check{c}}^{\hat{c}} \left(1 - e^{-y}\right) \quad dQ(y)}. \tag{3.29}$$

On the other hand the true vocabulary (scaled by sample size) converges almost surely in the LNRE regime to (cf. Equation (3.23))

$$\int_{\check{c}}^{\hat{c}} \frac{1}{y} \quad dQ(y). \tag{3.30}$$

We now claim that the Good-Turing vocabulary estimator *always strictly understimates*, i.e.,

$$\frac{\int_{\check{c}}^{\hat{c}} \frac{\left(1-e^{-y}\right)}{y} \quad dQ(y)}{\int_{\check{c}}^{\hat{c}} \left(1-e^{-y}\right) \quad dQ(y)} < \int_{\check{c}}^{\hat{c}} \frac{1}{y} \quad dQ(y), \tag{3.31}$$

whenever $dQ(\cdot)$ is not a single impulse distribution (corresponding to plain uniform LNRE distribution). By rearranging terms, we see that to prove Equation (3.31) it suffices to show that

$$\int_{\check{c}}^{\hat{c}} \frac{\left(1-e^{-y}\right)}{y} \quad dQ(y) < \left(\int_{\check{c}}^{\hat{c}} \left(1-e^{-y}\right) \quad dQ(y)\right) \left(\int_{\check{c}}^{\hat{c}} \frac{1}{y} \quad dQ(y)\right). \tag{3.32}$$

Since $dQ(\cdot)$ is a probability distribution, we can rewrite the desired inequality in Equation (3.32) as

$$\int_{\check{c}}^{\hat{c}} \int_{\check{c}}^{\hat{c}} \frac{\left(1-e^{-y_1}\right)}{y_1} \quad dQ(y_1)\, dQ(y_2) < \int_{\check{c}}^{\hat{c}} \int_{\check{c}}^{\hat{c}} \left(1-e^{-y_1}\right) \frac{1}{y_2} \quad dQ(y_1)\, dQ(y_2). \tag{3.33}$$

Combining the terms on both sides of Equation (3.33), the desired inequality becomes

$$\int_{\check{c}}^{\hat{c}} \int_{\check{c}}^{\hat{c}} \left(1-e^{-y_1}\right) \left(\frac{1}{y_2} - \frac{1}{y_1}\right) \quad dQ(y_1)\, dQ(y_2) > 0. \tag{3.34}$$

The integrand is zero whenever $y_1 = y_2$. Combining the integrands for the pairs (y_1, y_2) and (y_2, y_1) in Equation (3.34), the desired inequality becomes

$$\int_{\check{c}}^{\hat{c}} \int_{y_2}^{\hat{c}} \left(e^{-y_2} - e^{-y_1}\right) \left(\frac{1}{y_2} - \frac{1}{y_1}\right) \quad dQ(y_1)\, dQ(y_2) > 0. \tag{3.35}$$

This is readily true, since the integrand is strictly positive whenever $y_1 \neq y_2$:

$$\left(e^{-y_2} - e^{-y_1}\right) \left(\frac{1}{y_2} - \frac{1}{y_1}\right) > 0. \tag{3.36}$$

3.3.7 A Consistent Estimator of $\mathbb{O}_n$

We start with the limiting expression for scaled profiles in Equation (3.27) and come up with a natural estimator for $\mathbb{O}_n$. Our development leading to the estimator is somewhat heuristic and is aimed at motivating the structure of the estimator for the number of unseen words, $\mathbb{O}_n$. We formally state and prove its consistency at the end of this section.

A Heuristic Derivation

Starting from (3.27), let us first make the approximation that

$$\frac{k\varphi_k}{n} \approx \lambda_{k-1}, \quad k = 1, \ldots, n. \tag{3.37}$$

We now have the formal calculation

$$\begin{aligned}
\sum_{k=1}^{n} \frac{\varphi_k^n}{n} &\approx \sum_{k=1}^{n} \frac{\lambda_{k-1}}{k} && (3.38)\\
&= \sum_{k=1}^{n} \int_{\check{c}}^{\hat{c}} \frac{e^{-y} y^{k-1}}{k!}\, dQ(y) \\
&\approx \int_{\check{c}}^{\hat{c}} \frac{e^{-y}}{y} \left(\sum_{k=1}^{n} \frac{y^k}{k!} \right) dQ(y) && (3.39)\\
&\approx \int_{\check{c}}^{\hat{c}} \frac{e^{-y}}{y} \left(e^y - 1\right)\, dQ(y) && (3.40)\\
&\approx \frac{|\Omega_n|}{n} - \int_{\check{c}}^{\hat{c}} \frac{e^{-y}}{y}\, dQ(y). && (3.41)
\end{aligned}$$

Here the approximation in Equation (3.38) follows from the approximation in Equation (3.37); the approximation in Equation (3.39) involves swapping the outer discrete summation with integration and is justified formally later in the section; the approximation in Equation (3.40) follows because

$$\sum_{k=1}^{n} \frac{y^k}{k!} \to e^y - 1,$$

as $n \to \infty$; and the approximation in Equation (3.41) is justified from the convergence in Equation (3.23). Now, comparing Equation (3.41) with Equa-

tion (3.26), we arrive at an approximation for our quantity of interest:

$$\frac{\mathbb{O}_n}{n} \approx \int_{\check{c}}^{\hat{c}} \frac{e^{-y}}{y}\, dQ(y). \tag{3.42}$$

The geometric series allows us to write

$$\frac{1}{y} = \frac{1}{\hat{c}} \sum_{\ell=0}^{\infty} \left(1 - \frac{y}{\hat{c}}\right)^{\ell}, \quad \forall y \in (0, \hat{c})\,. \tag{3.43}$$

Approximating this infinite series by a finite summation, we have for all $y \in (\check{c}, \hat{c})$,

$$\begin{aligned} \frac{1}{y} - \frac{1}{\hat{c}} \sum_{\ell=0}^{M} \left(1 - \frac{y}{\hat{c}}\right)^{\ell} &= \frac{\left(1 - \frac{y}{\hat{c}}\right)^{M}}{y} \\ &\leq \frac{\left(1 - \frac{\check{c}}{\hat{c}}\right)^{M}}{\check{c}}. \end{aligned} \tag{3.44}$$

It helps to write the truncated geometric series as a power series in y:

$$\begin{aligned} & \frac{1}{\hat{c}} \sum_{\ell=0}^{M} \left(1 - \frac{y}{\hat{c}}\right)^{\ell} \\ = \; & \frac{1}{\hat{c}} \sum_{\ell=0}^{M} \sum_{k=0}^{\ell} \binom{\ell}{k} (-1)^k \left(\frac{y}{\hat{c}}\right)^k \\ = \; & \frac{1}{\hat{c}} \sum_{k=0}^{M} \left(\sum_{\ell=k}^{M} \binom{\ell}{k} \right) (-1)^k \left(\frac{y}{\hat{c}}\right)^k \\ = \; & \sum_{k=0}^{M} (-1)^k\, a_k^M y^k, \end{aligned} \tag{3.45}$$

where we have written

$$a_k^M := \frac{1}{\hat{c}^{k+1}} \left(\sum_{\ell=k}^{M} \binom{\ell}{k} \right). \tag{3.46}$$

Substituting the finite summation approximation in Equation (3.44) and its power series expression in Equation (3.45) into Equation (3.42) and swapping

the discrete summation with the integral, we can continue

$$\begin{aligned}\frac{\mathbb{O}_n}{n} &\approx \sum_{k=0}^{M} (-1)^k \, a_k^M \int_{\check{c}}^{\hat{c}} e^{-y} y^k \, dQ(y) \\ &= \sum_{k=0}^{M} (-1)^k \, a_k^M k! \lambda_k. \end{aligned} \tag{3.47}$$

Here, in Equation (3.47), we used the definition of λ_k from Equation (3.28). From the convergence in Equation (3.27), we finally arrive at our estimate:

$$\mathbb{O}_n \approx \sum_{k=0}^{M} (-1)^k \, a_k^M \, (k+1)! \;\; \varphi_{k+1}. \tag{3.48}$$

3.3.8 Consistency

Our main result is the demonstration of the consistency of the estimator guessed in Equation (3.48). Define

$$\hat{\mathbb{O}}_n \stackrel{\text{def}}{=} \sum_{k=0}^{M} (-1)^k \, a_k^M \, (k+1)! \;\; \varphi_{k+1}, \tag{3.49}$$

with a_k^M defined in Equation (3.46).

Theorem 1 *For any $\epsilon > 0$,*

$$\lim_{n \to \infty} \frac{\left| \mathbb{O}_n - \hat{\mathbb{O}}_n \right|}{n} \leq \epsilon$$

almost surely, as long as

$$M \geq \frac{\check{c} \log_2 e + \log_2 (\epsilon \check{c})}{\log_2 (\hat{c} - \check{c}) - 1 - \log_2 (\hat{c})}. \tag{3.50}$$

Proof: From Equation (3.26), we have

$$\begin{aligned}\frac{\mathbb{O}_n}{n} &= \frac{|\Omega_n|}{n} - \sum_{k=1}^{n} \frac{\varphi_k}{n} \\ &= \frac{|\Omega_n|}{n} - \sum_{k=1}^{n} \frac{\lambda_{k-1}}{k} - \\ &\quad \sum_{k=1}^{n} \frac{1}{k} \left(\frac{k\varphi_k}{n} - \lambda_{k-1} \right). \end{aligned} \tag{3.51}$$

The first term in the right–hand side (RHS) of Equation (3.51) converges as seen in Equation (3.23). The third term in the RHS of Equation (3.51) converges to zero, almost surely, as seen from Equation (3.27). The second term in the RHS of Equation (3.51), on the other hand,

$$\begin{aligned}\sum_{k=1}^{n} \frac{\lambda_{k-1}}{k} &= \int_{\check{c}}^{\hat{c}} \frac{e^{-y}}{y} \left(\sum_{k=1}^{n} \frac{y^k}{k!} \right) \, dQ(y) \\ &\to \int_{\check{c}}^{\hat{c}} \frac{e^{-y}}{y} \left(e^y - 1 \right) \, dQ(y), n \to \infty, \\ &= \int_{\check{c}}^{\hat{c}} \frac{1}{y} \, dQ(y) - \int_{\check{c}}^{\hat{c}} \frac{e^{-y}}{y} \, dQ(y). \end{aligned}$$

The monotone convergence theorem justifies the convergence in the second step above. Thus we conclude that

$$\lim_{n \to \infty} \frac{\mathbb{O}_n}{n} = \int_{\check{c}}^{\hat{c}} \frac{e^{-y}}{y} \, dQ(y) \tag{3.52}$$

almost surely. Coming to the estimator, we can write it as the sum of two terms:

$$\begin{aligned}&\sum_{k=0}^{M} (-1)^k \, a_k^M \, k! \lambda_k \\ &\quad + \sum_{k=0}^{M} (-1)^k \, a_k^M \, k! \left(\frac{(k+1)\, \varphi_{k+1}}{n} - \lambda_k \right). \end{aligned} \tag{3.53}$$

The second term in Equation (3.53) above is seen to converge to zero almost surely as $n \to \infty$, using Equation (3.27) and noting that M is a constant not depending on n. The first term in Equation (3.53) can be written, using the

definition of λ_k from Equation (3.28),

$$\int_{\check{c}}^{\hat{c}} e^{-y} \left(\sum_{k=0}^{M} (-1)^k \, a_k^M y^k \right) \, dQ(y). \tag{3.54}$$

Combining Equations (3.52) and (3.54), we have that, almost surely,

$$\lim_{n\to\infty} \frac{\mathbb{O}_n - \sum_{k=0}^{M} (-1)^k \, a_k^M \, (k+1)! \, \varphi_{k+1}}{n} = \int_{\check{c}}^{\hat{c}} e^{-y} \left(\frac{1}{y} - \sum_{k=0}^{M} (-1)^k \, a_k^M y^k \right) \, dQ(y). \tag{3.55}$$

Combining Equation (3.44) with Equation (3.45), we have

$$0 < \frac{1}{y} - \sum_{k=0}^{M} (-1)^k \, a_k^M y^k \leq \frac{\left(1 - \frac{\check{c}}{\hat{c}}\right)^M}{\check{c}}. \tag{3.56}$$

Using Equation (3.56), the quantity in Equation (3.55) can now be upper bounded by

$$\frac{e^{-\check{c}} \left(1 - \frac{\check{c}}{\hat{c}}\right)^M}{\check{c}}.$$

For M that satisfy Equation (3.50) this term is less than ϵ. The proof concludes.

3.4 Uniform Convergence and Rates of Convergence

Our main result has been the proposal of a non-parametric estimator that is consistent in the LNRE regime, with appropriately defined lower and upper constraints on the probabilities involved. These constraints (we used the terminology of $\hat{c}$ and $\check{c}$ for the constants, refer Equation (3.17)) were mostly for technical reasons (to enable the main proof in Section 3.3.8). Perhaps they are unnecessary and their removal would enlarge the technical scope of their result. Indeed, there is good precedence for this type of a general result: the main convergence result (cf. Equation (3.27)) used here from [74] does not impose such constraints. In this section we explore this direction in detail.

Once convergence of the estimator has been established, a natural question

to turn to is the rate of convergence. There are two parameters in the problem statement with which the rate of convergence can be evaluated. First is the number of terms in the estimator (denoted by M in the earlier section). The dependence of the error term with respect to M has been quantified explicitly in our main consistency result (cf. Theorem 1), while supposing the sample size is arbitrarily large. Second is the sample size itself (with either fixed M or letting it grow with the sample size, see the upcoming discussion in Section 3.4.1). In this section, we explore the rate of convergence of the estimator with respect to the sample size in the setting of a large deviation regime. While large deviation analysis is a classical topic in probability theory (cf. [77], [78]), it is particularly involved when applied to the sequence of random variables involving the scaled profiles. Using a recent result in the applied probability literature, we derive a large deviation result for the special case involving the uniform LNRE regime.

This result leads to a general conjecture on the large deviation rate of convergence for any LNRE regime. The conjecture is natural and has important ramifications. In particular, in many situations the available data allows for unseen element estimation of *subprocesses*. An instance involving natural language corpora is the following: with the English language we can automatically tag named entities and have their counts separately. One suspects that estimating the unseen elements of named entities separately (and the unseen elements in the non-named entities) should yield a better overall estimate than working with the whole monolithic data set. Indeed, as we will see shortly, the large deviation results support this natural inclination: the rate of convergence with respect to the monolithic process is slower than the average of the rates of convergence of the estimator working with the sub-processes separately.

We begin the material in this section with making the estimator independent of the upper and lower constraints on the probabilities as well as the number of terms involved in the summation.

3.4.1 Uniform Consistent Estimation

Consider the estimator for the unseen number of words (cf. Equation (3.49)):

$$\hat{\mathbb{O}}_n(M) \stackrel{\text{def}}{=} \sum_{k=0}^{M} (-1)^k \, a_k^M \, (k+1)! \;\; \varphi_{k+1}, \tag{3.57}$$

with the constants a_k^M defined in Equation (3.46) and reproduced here for convenience:

$$a_k^M := \frac{1}{\hat{c}^{k+1}} \left(\sum_{\ell=k}^{M} \left(\begin{array}{c} \ell \\ k \end{array} \right) \right). \tag{3.58}$$

This estimator is a function of M, the number of profile elements used in the process of estimation. An important issue with actually employing the estimator for the number of unseen elements (refer Equation (3.49)) is that it involves making a good choice of M knowing the parameter $\hat{c}$. In practice, there is no natural way to obtain any estimate on this parameter $\hat{c}$. It would be most useful if there were a way to modify the estimator in a way that it does not depend on the unobservable quantity $\hat{c}$. In this section we see that such a modification is possible, while still retaining the main theoretical performance result of consistency (cf. Theorem 1).

The first step to see the modification is in observing where the need for $\hat{c}$ arises: it is in writing the geometric series for the function $\frac{1}{y}$ (cf. Equations (3.43) and (3.44)). If we could let $\hat{c}$ along with the number of elements M itself depend on the sample size n, then we could still have the geometric series formula. More precisely, we have

$$\begin{aligned} \frac{1}{y} - \frac{1}{\hat{c}_n} \sum_{\ell=0}^{M_n} \left(1 - \frac{y}{\hat{c}_n}\right)^{\ell} &= \frac{1}{y} \left(1 - \frac{y}{\hat{c}_n}\right)^{M_n} \\ &\to \; 0, \quad n \to \infty, \end{aligned}$$

as long as

$$\frac{\hat{c}_n}{M_n} \to 0, \quad n \to \infty. \tag{3.59}$$

This simple calculation suggests that we can replace $\hat{c}$ and M in the formula for the estimator (cf. Equation (3.49)) by terms that depend on n and satisfy the condition expressed by Equation (3.59), thereby suppressing the dependence of the estimator on M (supposing that M and c_n are chosen as

increasing functions of n and satisfy Equation (3.59)):

$$\lim_{n\to\infty} \frac{|\mathbb{O}_n - \hat{\mathbb{O}}_n|}{n} = 0, \quad \text{almost surely.} \tag{3.60}$$

3.4.2 Convergence Rate Analysis

Our main convergence results have been in the limit of an arbitrarily large number of samples n. In practice, the sample size is always finite and it is of great utility to know how large a sample size suffices (in the sense of making the estimator reliable enough). A standard way to study such questions is to understand the rate of convergence of the normalized error (to zero) as a function of the sample size. In probability theory, there are two separate paths to study the convergence rate of the error in the estimate:

- *Large deviations*: In this regime, the event of interest is that the error stays a fixed (strictly non-zero) amount for all sample sizes. When the sample size is large, the event of interest in this regime is when the non-normalized error becomes arbitrarily large (leading to the term "large deviation"; here deviation is derived from the expected value of the error, which is zero in this case). Such an event is quite rare and, typically, the probability of this rare event converges exponentially fast to zero, as a function of the sample size. In such cases, the rate of convergence is measured as the normalized (by sample size) negative logarithm of the probability of the large deviation (this is the so-called *large deviation exponent*). The larger the exponent, the faster the rate of convergence of the estimator.

- *Central deviations*: In this regime, the event of interest is when the error in the estimate shrinks to zero along with the sample size. If the deviation of the error shrinks too fast, then that would be unrealistic to expect and the probability of such an event would converge to 0. On the other hand, if the deviation of the error shrinks too slowly, then this event would be very much expected and the probability of the corresponding event would converge to 1. Typically, the "correct" rate at which the error should shrink to zero is $\frac{1}{\sqrt{n}}$: in this case, the probability of such an event would then be nontrivial (strictly between 0

and 1). The result that characterizes the limiting probability is usually called a *central limit theorem*.

We cover the former scenario in detail next, with varying degrees of success.

3.4.3 Large Deviation Analysis

The focus of this section is in understanding the behavior of error in the convergence of the estimator (cf. Equation (3.57)):

$$\mathbb{P}\left[|\hat{\mathbb{O}}_n - \sum_{k=0}^{M} (-1)^k \, a_k^M \, (k+1)! \;\; \lambda_k| \geq n\epsilon\right] \tag{3.61}$$

as a function of n for a fixed $\epsilon > 0$. This regime is said to involve "large deviations," since the error is allowed to scale linearly with n (and thus be as large as the estimates themselves (up to a constant factor)), as the sample size grows. We do know from the main result of [74] (cf. Equation (3.27)) that the sequence of probabilities in Equation (3.61) goes to zero as n grows to infinity. The key question of interest is the *rate* of convergence to zero. We see that the basic underlying reason for our main result is the convergence of *scaled profiles*: from Section 3.3.5 (and Equation (3.27) in particular) the key underlying convergence result is

$$\sum_{k=1}^{n} \left| \frac{k\varphi_k}{n} - \lambda_{k-1} \right| \longrightarrow 0, \quad \text{almost surely,} \tag{3.62}$$

as $n \to \infty$. To understand the convergence rate of the probabilities in Equation (3.61) it would help greatly if the convergence rate of the underlying result (cf. Equation (3.62)) were known. To see the form of the desired convergence rate, we introduce some notation. Denote the infinite dimensional random vector

$$\mathbf{\Phi}_n \stackrel{\text{def}}{=} \left(\frac{\varphi_1}{n}, \frac{2\varphi_2}{n}, \ldots, \frac{n\varphi_n}{n}, 0, 0, \ldots \right). \tag{3.63}$$

Denote the infinite dimensional deterministic vector

$$\mathbf{\Lambda} \stackrel{\text{def}}{=} (\lambda_0, \lambda_1, \lambda_2, \ldots,). \tag{3.64}$$

Then the convergence result in Equation (3.62) can be rewritten in the new notation as

$$\|\mathbf{\Phi}_n - \mathbf{\Lambda}\|_1 \to 0, \quad \text{almost surely,} \tag{3.65}$$

as $n \to \infty$. Here we have used the standard notation of the L_1 norm: for any infinite dimensional vector $\mathbf{x}$

$$\|\mathbf{x}\|_1 \stackrel{\text{def}}{=} \sum_{k=1}^{\infty} |x_k|. \tag{3.66}$$

Now consider the large deviation event

$$\mathbf{\Phi}_n = \mathbf{\Theta} \neq \mathbf{\Lambda}. \tag{3.67}$$

Here $\mathbf{\Theta}$ is a deterministic infinite dimensional probability vector. Suppose the probability of this large deviation event decays to zero as follows:

$$\mathbb{P}\left[\mathbf{\Phi}_n = \mathbf{\Theta} \neq \mathbf{\Lambda}\right] \doteq e^{-nf(\mathbf{\Theta},\mathbf{\Lambda})}. \tag{3.68}$$

Here we used standard notation of $\doteq$ from the large deviation literature: for any sequence of numbers $\{b_n\}_n$

$$b_n \doteq e^{-na} \tag{3.69}$$

is equivalent to the following limiting statement:

$$\lim_{n\to\infty} \frac{-\log_e(b_n)}{n} = a. \tag{3.70}$$

Intuitively, the statement in Equation (3.68) says that $f(\mathbf{\Theta}, \mathbf{\Lambda})$ is the (exponential) decay rate of the probability of the large deviation event in Equation (3.67). The larger the value of $(\mathbf{\Theta}, \mathbf{\Lambda})$, the faster the convergence rate. If a result of the type in Equation (3.68) is known, then it can be used to deduce the convergence rate in Equation (3.61). This is done in the following way. First we see that both the true value and the estimated value are the

same linear functional of the $\mathbf{\Lambda}$ and $\mathbf{\Phi}$ vectors, respectively:

$$\hat{\mathbb{O}}_n = \mathbf{a}_n^T \mathbf{\Phi}_n \tag{3.71}$$

$$\sum_{k=0}^{M} (-1)^k \, a_k^M \, (k+1)! \;\; \lambda_k = \mathbf{a}_n^T \mathbf{\Lambda}_n. \tag{3.72}$$

So, we can rewrite the large deviation event of interest (cf. Equation (3.61)) as

$$|\hat{\mathbb{O}}_n - \sum_{k=0}^{M} (-1)^k \, a_k^M \, (k+1)! \;\; \lambda_k| \geq n\epsilon = \bigcup_{\{\mathbf{\Theta}: |\mathbf{a}_n^T (\mathbf{\Theta} - \mathbf{\Lambda})| \geq n\epsilon\}} \{\mathbf{\Phi}_n = \mathbf{\Theta}\}. \tag{3.73}$$

So, the corresponding large deviation probability is (from Equation (3.68))

$$\mathbb{P}\left[\frac{|\hat{\mathbb{O}}_n - \sum_{k=0}^{M} (-1)^k \, a_k^M \, (k+1)! \;\; \lambda_k|}{n} \geq \epsilon \right] \doteq e^{-\min_{\Theta: |\mathbf{a}_n^T (\mathbf{\Theta} - \mathbf{\Lambda})| \geq n\epsilon} f(\mathbf{\Theta}, \mathbf{\Lambda})}. \tag{3.74}$$

Thus having a formula for the large deviation exponent $f(\mathbf{\Theta}, \mathbf{\Lambda})$ in Equation (3.68) directly leads to the large deviation exponent for the convergence error of the estimator. Finding such a formula is the focus of the next two sections.

Large Deviation of the Empirical Distribution of IID Samples

Large deviation analysis is a classical topic in probability theory ([77], [78]). We start with a classical result in this area to help set the tone for our context. Consider a sequence of i.i.d. random variables $X_1, \ldots, X_n$ with common probability distribution $\mathbf{P}_X$. Consider the empirical probability distribution $\hat{\mathbf{P}}_n$ defined in the usual way:

$$\hat{\mathbf{P}}_n(x) \stackrel{\text{def}}{=} \frac{\sum_{k=1}^{n} \mathbf{1}_{\{X_k = x\}}}{n} \tag{3.75}$$

where $\mathbf{1}_A$ is the indicator function — it is 1 when the event A occurs and zero otherwise. The functional strong law of large numbers asserts that

$$\|\hat{\mathbf{P}}_n - \mathbf{P}_X\|_1 \to 0, \text{almost surely} \tag{3.76}$$

as $n \to \infty$. The corresponding large deviation event is

$$\hat{\mathbf{P}}_n = \mathbf{Q} \neq \mathbf{P}_X. \tag{3.77}$$

The *Sanov theorem* characterizes the large deviation exponent:

$$\mathbb{P}\left[\hat{\mathbf{P}}_n = \mathbf{Q}\right] \doteq e^{-nD(\mathbf{Q}\|\mathbf{P}_X)}. \tag{3.78}$$

Here $D(\cdot\|\cdot)$ is the *relative entropy* between the probability measures used in the argument:

$$D(\mathbf{Q}\|\mathbf{P}_X) \stackrel{\text{def}}{=} \sum_x \mathbf{Q}(x) \log_e \frac{\mathbf{Q}(x)}{\mathbf{P}_X(x)}. \tag{3.79}$$

3.4.4 Large Deviation of the Scaled Profiles

While the classical result of Sanov is for empirical distribution of a sequence of i.i.d. random variables, our interest here is with the sequence of scaled profiles $\mathbf{\Phi}_n$. We do have the basic convergence result (cf. Equation (3.27))

$$\|\mathbf{\Phi}_n - \mathbf{\Lambda}\|_1 \to 0, \quad \text{almost surely} \tag{3.80}$$

as $n \to \infty$. However the scaled profile vector $\mathbf{\Phi}_n$ is not the empirical distribution of an i.i.d. sequence. Probabilistically, it is a *Markov* chain:

$$\mathbb{P}\left[\mathbf{\Phi}_n = \mathbf{\Theta}_n | \mathbf{\Phi}_k = \mathbf{\Theta}_k, k = 1 \ldots, n-1\right] = \mathbb{P}\left[\mathbf{\Phi}_n = \mathbf{\Theta}_n | \mathbf{\Phi}_{n-1} = \mathbf{\Theta}_{n-1}\right]. \tag{3.81}$$

Furthermore, it is a homogeneous Markov chain in the sense that the transition probabilities between successive states do not depend on the sample size n. The large deviation analysis for such a Markov chain is complicated and is the main topic of [79] for the special case when the LNRE regime has all equal probabilities (i.e., the shadow is an impulse $\delta(\cdot)$ function). The remarkable result of [79] (in particular, Theorem 2.5 in [79]) is that the large deviation exponent of the scaled profiles behaves in much the same way as prescribed by Sanov's theorem for the empirical distributions of an i.i.d. process. Put another way (rephrasing Theorem 2.5 of [79]),

$$\mathbb{P}\left[\mathbf{\Phi}_n = \mathbf{\Theta}\right] \doteq e^{-nD(\mathbf{\Theta}\|\mathbf{\Lambda})}. \tag{3.82}$$

In the notation of Equation (3.68), we have

$$f(\mathbf{\Theta}, \mathbf{\Lambda}) = D(\mathbf{\Theta}||\mathbf{\Lambda}). \tag{3.83}$$

The proof of this result is very technical, but the gist of the argument appears to approximate the vector of scaled profiles by an appropriate sequence of empirical distributions of an i.i.d. process and then invoke the Sanov theorem. We conjecture that the result of [79] (as stated in Equation (3.82)) holds for more general LNRE distributions than just uniform probability case. This task appears to be challenging from a technical perspective, but it has very interesting ramifications as we will see next.

3.4.5 Quantitative Properties of the Large Deviation Exponent

Consider two LNRE regimes (with corresponding limits $\mathbf{\Lambda}_A$ and $\mathbf{\Lambda}_B$) and their *mixture*: $\mu\mathbf{\Lambda}_A + (1-\mu)\mathbf{\Lambda}_B$. One would expect that the estimator working directly on the mixture of the LNRE regimes (a single monolithic process) would have a *slower* rate of convergence than the mixture of the convergence rates of the individual LNRE regimes separately. Indeed, this is justified mathematically:

Proposition 2 *For any* $\mu \in (0,1)$,

$$D(\mu\mathbf{\Theta}_A + (1-\mu)\mathbf{\Theta}_B \| \mu\mathbf{\Lambda}_A + (1-\mu)\mathbf{\Lambda}_B) \leq \mu D(\mathbf{\Theta}_A \| \mathbf{\Lambda}_A) + D(\mathbf{\Theta}_B \| \mathbf{\Lambda}_B). \tag{3.84}$$

This follows directly from the convexity property of the relative entropy function (refer Example 3.19 in [80]).

3.4.6 Large Deviations for Estimator with Finite Number of Terms

Our limiting result came in two varieties (one with finite number M of profile terms, and the other with letting M_n scale to infinity with the sample size n). The large deviation result of the previous section was for the latter case.

In this section, we borrow the results for finite number M of profile terms in [79] to derive the large deviation exponent for the unseen element estimator.

The key step is the replacement of the large deviation exponent in Equation (3.82) for a finite (and fixed) value of M. We first define $\mathbf{\Phi}_n^M$ as a finite dimensional random vector: it consists of the first $M+1$ terms of the random vector $\mathbf{\Phi}_n$. Analogously, we define $\mathbf{\Lambda}^M$ as a finite dimensional deterministic vector: it consists of the first $M+1$ terms of the deterministic vector $\mathbf{\Lambda}$. We do have the limiting result

$$\|\mathbf{\Phi}_n^M - \mathbf{\Lambda}^M\|_1 \to 0, \quad \text{almost surely} \tag{3.85}$$

as $n \to \infty$. Analogous to the discussion in the case when we had arbitrarily large M, we ask for the large deviation exponent: $f_M(\mathbf{\Theta}^M, \mathbf{\Lambda}^M)$, defined as

$$\mathbb{P}\left[\mathbf{\Phi}_n^M = \mathbf{\Theta}^M\right] \doteq e^{-n f_M\left(\mathbf{\Theta}^M, \mathbf{\Lambda}^M\right)}. \tag{3.86}$$

If we have an expression for $f_M(\cdot,\cdot)$, then we can use it to derive the large deviation exponent for the unseen element estimator just as in Equation (3.74):

$$\mathbb{P}\left[\frac{|\hat{\mathbb{O}}_n^M - \sum_{k=0}^M (-1)^k a_k^M (k+1)!\ \lambda_k|}{n} \geq \epsilon\right] \doteq e^{-\min_{\Theta:|\mathbf{a}_n^T(\mathbf{\Theta}-\mathbf{\Lambda})|\geq n\epsilon} f(\mathbf{\Theta},\mathbf{\Lambda}).} \tag{3.87}$$

We start with a uniform LNRE distribution with the probability of each word being $\frac{\beta}{n}$. Then the limit of the scaled profiles denoted by $\mathbf{\Lambda}_\beta$ is particularly simple: the k^{th} element of this vector is

$$\lambda_k = e^{-\beta}\frac{\beta^{k-1}}{k-1!} \tag{3.88}$$

From Theorem 2.5 of [79], the large deviation exponent $f_M(\cdot,\cdot)$ is calculated as follows:

$$\min_{\mathbf{\Theta}:\text{first } M+1 \text{ terms are same as that of } \mathbf{\Theta}^M} D\left(\mathbf{\Theta}\|\mathbf{\Lambda}\right). \tag{3.89}$$

The minimizing argument $\mathbf{\Theta}^*$ is unique and and can be computed explicitly. Using Lagrange multipliers it is shown in Section 2 of [79] that the solution takes the form

$$\mathbf{\Theta}^*(k) = C\mathbf{\Lambda}_{\rho\beta}, \quad \forall k \geq M+1. \tag{3.90}$$

Here $\rho, C \geq 0$ are parameters to be optimized over. Further $\Lambda_{\rho\beta}$ is defined as the Poisson random vector with $\rho\beta$ as the height of a uniform LNRE distribution that would yield the Poisson limit $\Lambda_{\rho\beta}$ where β is the original height of the uniform LNRE distribution under study. The authors in [79] refer to C, β as the "twist" parameters. Here ρ is related to the Lagrange multiplier for the conservation constraint

$$\sum_{k=0}^{M} k\mathbf{\Phi}(k) = \beta$$

while C is a normalization constant ensuring that

$$\sum_{k=0}^{\infty} \mathbf{\Phi}(k) = 1.$$

We define ρ to be the unique positive root of the equation

$$\frac{\rho\beta - \sum_{k=0}^{M} k\mathbf{\Lambda}_{\rho\beta}(k)}{1 - \sum_{k=0}^{M} \mathbf{\Lambda}_{\rho\beta}(k)} = \frac{\beta - \sum_{k=0}^{M} k\mathbf{\Theta}(k)}{1 - \sum_{k=0}^{M} \Theta(k)}. \tag{3.91}$$

This equation is shown in [79] to have a unique solution for C as

$$C^* = \frac{1 - \sum_{k=0}^{M} \mathbf{\Theta}(k)}{1 - \sum_{k=0}^{M} \mathbf{\Lambda}_{\rho\beta}(k)} = \frac{\beta - \sum_{k=0}^{M} k\mathbf{\Theta}_{\rho\beta}(k)}{\rho\beta - \sum_{k=0}^{M} k\mathbf{\Lambda}_{\rho\beta}(k)}. \tag{3.92}$$

3.5 Experiments

Having seen the theoretical properties of the estimator, we now look at its empirical performance in estimating the number of unseen elements. A natural setting in which this can be studied is by looking at natural language corpora which are intrinsically endowed with the large number of rare events property. Based on [37, Section 2.4], we assume that portions (such as represented using 10%, 20%, 30%, 40% and 50%) of the corpus are located in the LNRE zone, the range of vocabulary sizes where the vocabulary size is increasing linearly with word tokens. This assumption is plausible owing to the significant number of words that are occurring once.

3.5.1 Corpora

In our experiments we used the following corpora:

1. The *British National Corpus* (BNC): A corpus of about 100 million words of written and spoken British English from the years 1975-1994.

2. The *New York Times Corpus* (NYT): A corpus of about 5 million words.

3. The *Malayalam Corpus* (MAL): A collection of about 2.5 million words from varied articles in the Malayalam language from the Central Institute of Indian Languages.

4. The *Hindi Corpus* (HIN): A collection of about 3 million words from varied articles in the Hindi language also from the Central Institute of Indian Languages.

3.5.2 Methodology

We would like to see how well our estimator performs in terms of estimating the number of unseen elements. A natural way to study this is to use only half of an existing corpus to be observed and estimate the number of unseen elements (assuming the the actual corpus is twice the observed size). While our estimator has no assumptions on the relative sizes of the sample in comparison with that of the population, without loss of generality, we can consider that the population is twice the size of the sample. This ensures that we are operating in a region where vocabulary grows linearly with sample size. We then check numerically how well our estimator performs with respect to the "true" value. We use a subset (the first 10%, 20%, 30%, 40% and 50%) of the corpus as the *observed sample* to estimate the vocabulary over twice the sample size.

While the notions of a *word* and hence *vocabulary* depend on how a natural language corpus is split into observations (or tokens), we consider the simplest form of tokenization—a token is a string of letters that is delimited by space. This disregards the productive facility of a language and places the corpus of a productive language such as Malayalam on par with that of English. This is justified here since our objective is only to show the performance of the

estimator. We also clean the corpora by removing punctuation, numbers and other non-textual characters. For the English corpora we used a set of stop words comprising the high frequency function words in the language. No such filtering was used for Malayalam and Hindi corpora. Since we assume an i.i.d sample of observations in the formulation of the estimator, we consider the sample as a bag of words and disregard any underlying linguistic structure.

We compare the following parametric and nonparametric estimators.

Nonparametric: Along with our proposed estimator (in Section 3.3), the following canonical estimators (discussed in Chapter 2 and available in [27] and [37]) are studied.

1. Our proposed estimator $\mathbb{O}_n$: Since the estimator is rather involved we consider only small values of M (we see empirically that the estimator converges for very small values of M itself) and choose $\hat{c} = M$. This allows our estimator for the number of unseen elements to be of the following form, for different values of M:

M	$\mathbb{O}_n$
1	$2(\varphi_1 - \varphi_2)$
2	$\frac{3}{2}(\varphi_1 - \varphi_2) + \frac{3}{4}\varphi_3$
3	$\frac{4}{3}(\varphi_1 - \varphi_2) + \frac{8}{9}\left(\varphi_3 - \frac{\varphi_4}{3}\right)$

   Using this, the estimator of the true vocabulary size is simply

$$\mathbb{O}_n + V. \tag{3.93}$$

   Here (cf. Equation (3.25))

$$V = \sum_{k=1}^{n} \varphi_k^n. \tag{3.94}$$

   In the simulations below, we have considered M large enough until we see numerical convergence of the estimators: in all the cases, no more than a value of 4 is needed for M. For the English corpora, very small values of M suffice — in particular, we have considered the average of the first three different estimators (corresponding to the first three values of M). For the non-English corpora, we have needed to consider $M = 4$.

2. Gandolfi-Sastri estimator,

$$V_{\mathrm{GS}} \stackrel{\mathrm{def}}{=} \frac{n}{n-\varphi_1}\left(V+\varphi_1\gamma^2\right), \tag{3.95}$$

where

$$\gamma^2 = \frac{\varphi_1 - n - V}{2n} + \frac{\sqrt{5n^2+2n(V-3\varphi_1)+(V-\varphi_1)^2}}{2n}.$$

3. Chao estimator,

$$V_{\mathrm{Chao}} \stackrel{\mathrm{def}}{=} V + \frac{\varphi_1^2}{2\varphi_2}. \tag{3.96}$$

4. Good-Turing estimator,

$$V_{\mathrm{GT}} \stackrel{\mathrm{def}}{=} \frac{V}{\left(1-\frac{\varphi_1}{n}\right)}. \tag{3.97}$$

5. "Simplistic" estimator,

$$V_{\mathrm{Smpl}} \stackrel{\mathrm{def}}{=} V\left(\frac{n_{\mathrm{new}}}{n}\right). \tag{3.98}$$

Here the supposition is that the vocabulary size scales linearly with the sample size (here n_{new} is the new sample size).

6. Extrapolated estimator,

$$V_{\mathrm{Ext}} \stackrel{\mathrm{def}}{=} V + \left(\frac{\varphi_1}{n}\right) n_{new}. \tag{3.99}$$

Here the supposition is that the vocabulary growth rate at the observed sample size is given by the ratio of the number of *hapax legomena* (number of "singletons," φ_1 in our notation) to the sample size (cf. [37] pp. 50).

Parametric: In this study we consider the state-of-the-art parametric estimators, as surveyed by [38]. For the purpose of comparison we use the estimator based on the Zipf-Mandelbrot law (ZM) as well as its finite version (fZM). We are aided in this study by the availability of the implementations provided in the `ZipfR` package and their default settings.

3.6 Results and Discussion

The performance of the different estimators as percentage errors of the true vocabulary size using different corpora are tabulated in Tables 3.2-3.5. The Figures 3.2-3.5 compare the performance of our estimator with the best estimators (nonparametric and parametric).

We now summarize some important observations based on these tables and the figures.

- We see that our estimator compares quite favorably with the best of the state of the art estimators. The best state-of-the-art estimator is a parametric one (ZM), while ours is nonparametric.

- From Table 3.2 and Table 3.3 as well from Figures 3.2 and 3.3, we see that our estimate is quite close to the true vocabulary, at all sample sizes.

- Again, on the two non-English corpora (refer Tables 3.4 and 3.5 for actual values), we see that our estimator compares favorably with the best estimator of vocabulary size and at some sample sizes even surpasses it. This is represented in the Figures 3.5 and 3.4.

- On the BNC and Hindi corpora, our estimator outperforms the others. On the NYT corpus, the ZM-based estimator outperforms the others. On the Malayalam corpus, the Gandolfi-Sastri estimator and the ZM-based estimator outperform other estimators.

- We also observe that the performance of the various estimators seems to be corpus dependent. For instance, in the English corpora, the performance of the Chao estimator is better than that of the Gandolfi-Sastri estimator. For the non-English corpora, however, the relative performance of the same two estimators is reversed.

- The nonparametric Good-Turing estimator widely underestimates the vocabulary; this is true in each of the four corpora studied and at all sample sizes. This bears out the result that we show in Section 3.3.6, which is the tendency to underestimate vocabulary sizes owing to the assumption of uniform LNRE. While we showed theoretically that the

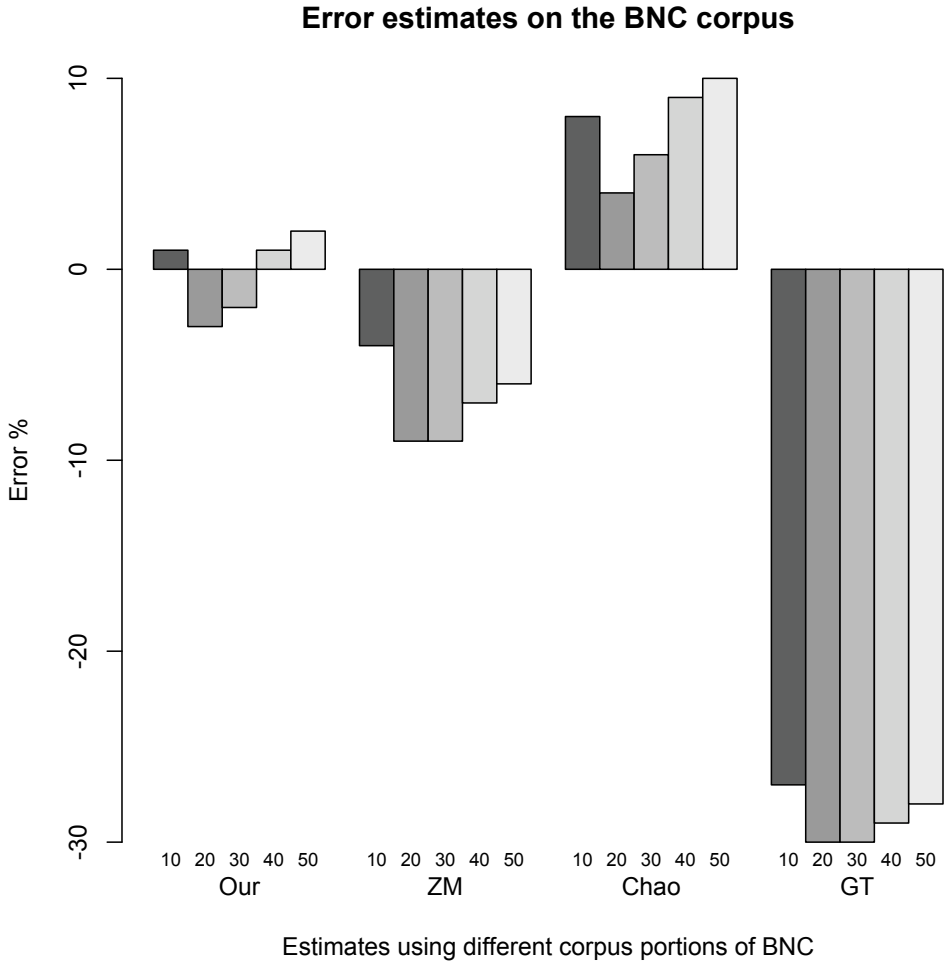

Figure 3.2: Comparison of estimation error of the best estimators with the Good-Turing estimator and our proposed estimator on the BNC corpus. Our estimator *outperforms* ZM. Good-Turing estimator widely *underestimates* vocabulary size.

Table 3.2: Comparison of estimates of vocabulary size for the BNC corpus as percentage errors w.r.t. the true value. A negative value indicates an underestimate. Our estimator *outperforms* the other estimators at all sample sizes.

Sample	True	% error w.r.t. the true value							
(% of corpus)	value	Our	GT	ZM	fZM	Smpl	Ext	Chao	GS
10	153912	1	-27	-4	-8	46	23	8	-11
20	220847	-3	-30	-9	-12	39	19	4	-15
30	265813	-2	-30	-9	-11	39	20	6	-15
40	310351	1	-29	-7	-9	42	23	9	-13
50	340890	2	-28	-6	-8	43	24	10	-12

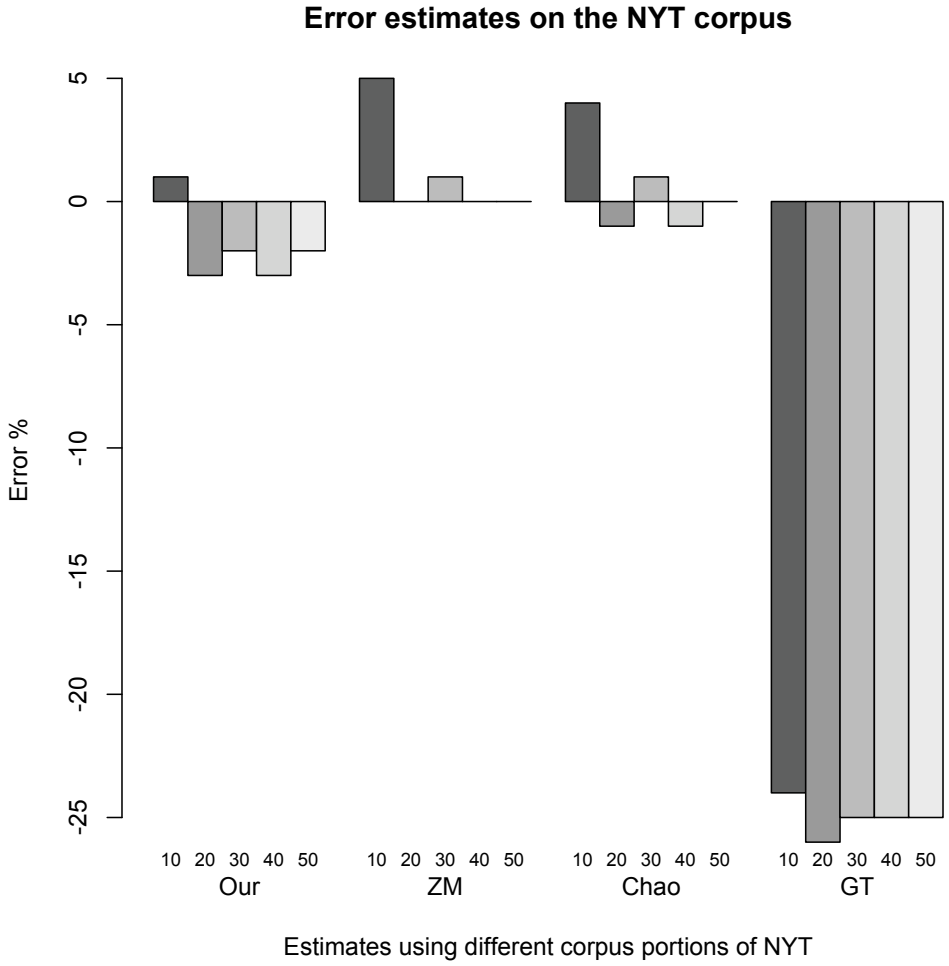

Figure 3.3: Comparison of estimation error of the best estimators with the Good-Turing estimator and our proposed estimator on the NYT corpus. Our estimator *compares favorably* with ZM and Chao. Our estimator *outperforms* ZM. Good-Turing estimator widely *underestimates* vocabulary size.

Table 3.3: Comparison of estimates of vocabulary size for the NYT corpus as percentage errors w.r.t. the true value. A negative value indicates an underestimate. Our estimator *compares favorably* with ZM and Chao.

Sample	True	% error w.r.t. the true value							
(% of corpus)	value	Our	GT	ZM	fZM	Smpl	Ext	Chao	GS
10	37346	1	-24	5	-8	48	28	4	-8
20	51200	-3	-26	0	-11	46	22	-1	-11
30	60829	-2	-25	1	-10	48	23	1	-10
40	68774	-3	-25	0	-10	49	21	-1	-11
50	75526	-2	-25	0	-10	50	21	0	-10

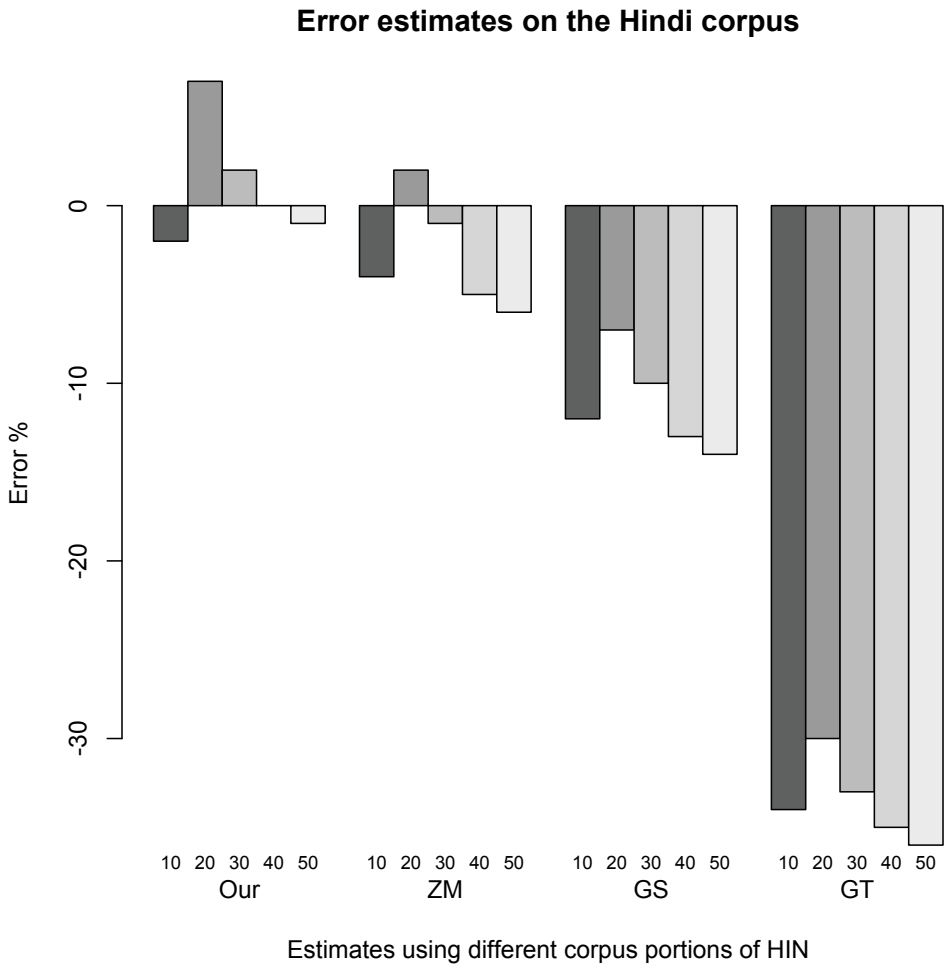

Figure 3.4: Comparison of estimation error of the best estimators with the Good-Turing estimator and our proposed estimator on the Hindi corpus. Our estimator *outperforms* the other estimators at certain sample sizes.

Table 3.4: Comparison of estimates of vocabulary size for the Hindi corpus as percentage errors w.r.t. the true value. A negative value indicates an underestimate. Our estimator *outperforms* the other estimators at certain sample sizes.

Sample	True	% error w.r.t. the true value							
(% of corpus)	value	Our	GT	ZM	fZM	Smpl	Ext	Chao	GS
10	47639	-2	-34	-4	-9	25	32	31	-12
20	71320	7	-30	2	-1	34	43	51	-7
30	93259	2	-33	-1	-5	30	38	42	-10
40	113186	0	-35	-5	-7	26	34	39	-13
50	131715	-1	-36	-6	-8	24	33	40	-14

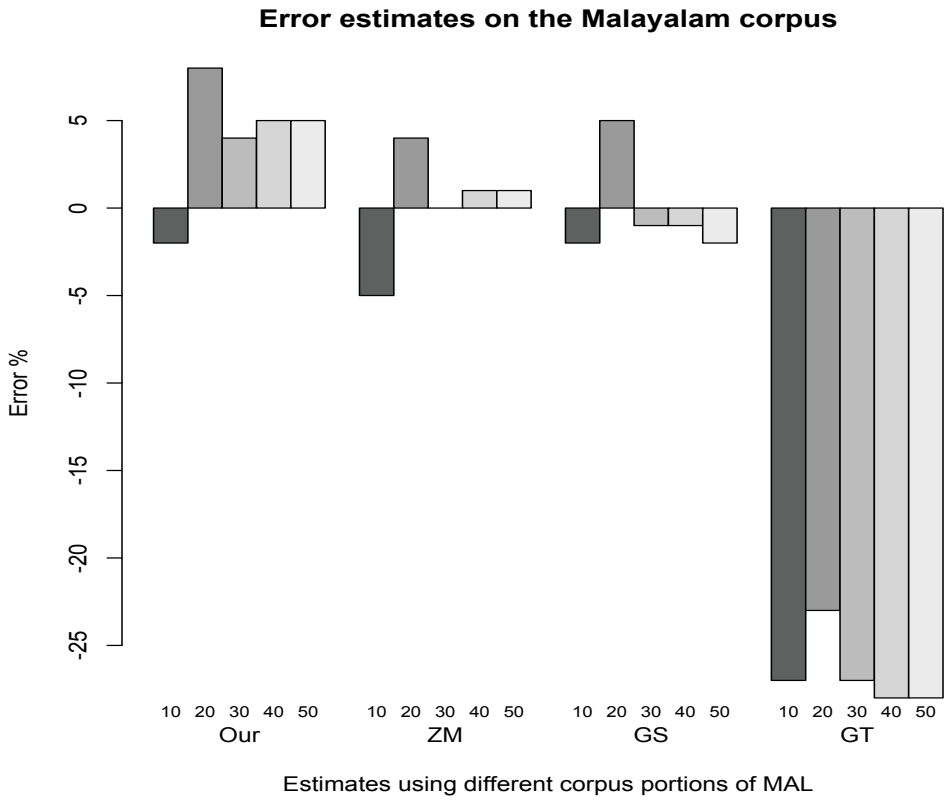

Figure 3.5: Comparison of estimation error of the best estimators with the Good-Turing estimator and our proposed estimator on the Malayalam corpus. Our estimator *compares favorably* with ZM and GS.

Table 3.5: Comparison of estimates of vocabulary size for the Malayalam corpus as percentage errors w.r.t. the true value. A negative value indicates an underestimate. Our estimator *compares favorably* with ZM and GS.

Sample (% of corpus)	True value	% error w.r.t. the true value							
		Our	GT	ZM	fZM	Smpl	Ext	Chao	GS
10	146547	-2	-27	-5	-10	9	34	82	-2
20	246723	8	-23	4	-2	19	47	105	5
30	339196	4	-27	0	-5	16	42	93	-1
40	422010	5	-28	1	-4	17	43	95	-1
50	500166	5	-28	1	-4	18	44	94	-2

uniform LNRE assumption leads to an underestimation, results of simulation show that the degree of underestimation is of the order of 25–30%.

3.7 Discussion

We showed in Section 3.2.2 that, under the assumption of uniform LNRE (all low probability events considered to be equally likely), the Good-Turing vocabulary size estimator is consistent. Furthermore, in Section 3.3.6 we showed that the estimator is not consistent when we the underlying distribution departs from the uniform LNRE regime and that it underestimates vocabulary size (even in the simple case of a binary mixture of uniform LNRE). Considering the fact that this estimator has a wide applicability, we note this serious limitation for use with natural language corpora. This is because we know that rare words are not just rare but are *unequally* rare.

To remedy this situation we propose a new estimator of vocabulary size that takes into account the non-uniform LNRE property of word frequency distributions, and we show that it is statistically consistent. Following the taxonomy of Bunge and Fitzpatrick [26], it is possible to classify our estimator as belonging to the infinite population regime with a multinomial sample where scaled frequencies are seen as converging to a continuous mixture of Poisson distributions. Rather than proceeding to find the best parametric fit to the observed data, we instead seek constants that approximate the continuous mixture of Poisson distributions by a discrete mixture. We then use the constants and the frequency counts to estimate the unseen vocabulary size. Thus, in essence, our estimator can be considered as being a nonparametric estimator of vocabulary, with an underlying parametric model that is transparent to the estimation process.

We then perform a large deviations analysis to understand the behavior of error in the convergence of the estimator. We then show that theoretically one can expect that the estimator working directly on the mixture of the LNRE regimes (a single monolithic process) would have a *slower* rate of convergence than the mixture of the convergence rates of the individual LNRE regimes separately. The implications of this result could be construed to be that if one were to find two subprocesses of the main generative process of

rare events, then the estimation procedure of vocabulary size of the main process would converge faster via estimation of vocabulary sizes in the two subprocesses. This of course requires that the two subprocesses in turn be in the LNRE zone. We attempted to show this empirically, but the practical difficulty of identifying subprocesses of the underlying rare event generative process (which behave differently as compared with the main process) came as a major impediment.

Comparing the performance of the proposed estimator with that of the state-of-the-art estimators on large corpora from different languages, we see that the performance of our estimator is comparable to that of the state of the art parametric and nonparametric estimators. While we notice a strong dependence of estimator performance with the underlying distribution of rare events (and hence the language of the corpus), we do not see a single estimator as emerging to be the best overall in terms of performance. This observation has been pointed out in the review papers by Bunge and Fitzpatrick as well as by Gandolfi and Sastri. However, comparing the performance of our estimator with that of the Good-Turing estimator, we see that our estimator by far outperforms its performance. Similarly, in comparison with Chao's estimator that is regarded to have a more general applicability than the others [26], our estimator shows better performance in certain corpora. The same is true of the Gandolfi-Sastri estimator which has been claimed to be the best nonparametric estimator of vocabulary size in a nonuniform regime. Again, our estimator shows better performance when compared with the state-of-the-art parametric estimator that is based on the Zipf-Mandelbrot law on some corpora.

That no estimator emerges as being the best overall seems to suggest that the underlying generative process of word occurrence influences the choice of estimator of vocabulary size. We leave it to future work to further characterize the underlying process and then relate this with the choice of a suitable estimator.

CHAPTER 4

AUTOMATIC FLUENCY ASSESSMENT

The purposes of the study described in this chapter are highlighted below.

1. While the effects of temporal aspects of speech on perceptions of fluency have been well understood (as mentioned in Section 2), the effects of qualitative aspects of speech are less understood. In this study we attempt to address this lesser known aspect of fluency—the extent to which effectiveness of speech production as a function of a productive vocabulary influences perceptions of fluency. Here we choose lexical richness as measured from the utterances to be indicative of the person's productive vocabulary.

2. We would like to make signal level measurements leading to a set of quantifiers of temporal aspects of speech production and verify that they quantify human perceptions of fluency reasonably well as evidenced in previous studies.

3. We would like to explore the possibility of designing alternate systems of automatic fluency assessment for spontaneous speech that are less reliant on ASR for better suitability to resource-scarce scenarios.

4. Drawing insights from studies in social psychology, we would like to find out the extent to which automatic assessments using a *thin-slice* of the original utterance agree with those made using the entire utterance.

This chapter is organized as follows. In Section 4.1 we describe the data set that we used for our experiments. In Section 4.2 we discuss the quantifiers of fluency covering both qualitative as well as quantitative aspects of speech. Here we consider the extent to which various quantifiers of fluency influence human perceptions of fluency (represented by the human-assigned fluency scores). This sets the stage for the choice of quantifiers in the design of

the automatic fluency assessment system, discussed in Section 4.3. Here we discuss the system architecture and its performance on the chosen data set. An important aspect of automatic assessment is the *thin-slice* assessment, which is based on a random snippet of the entire utterance. In Section 4.4 we show that automatic thin-slice assessment is not significantly different from that based on the entire utterance at the 5% level. Finally, we discuss our experiments and related observations in the context of the available studies in Section 4.5.

4.1 Data

For the purpose of our experiments we used the rated speech corpus of second language English learners constructed by the UIUC Speech and Language Engineering Group [81]. This corpus is a collection of spontaneous speech (and the corresponding transcription) from 28 speakers representing six language backgrounds and five proficiency levels.

The speech was recorded in a sound-attenuated setting and was collected using prompts consisting of eight questions using the format of the TOEFL iBT and that of the SPEAK test. Of these, two questions required the participant to describe a movie that they liked and a country they wanted to visit. Two questions involved describing a picture and two others required the speakers to give their opinion on a social issue after reading a short passage. Finally, there were two questions asking the speakers for directions based on a map.

Based on the speaking rubrics of TOEFL internet-based test (iBT, refer to Table 4.1 for the general description of proficiency at each score level) the utterances were rated for fluency on a 0-4 point scale (with 0.5 increments added to get a refined picture of the variations; 0 indicating no response and 4 indicating native-like fluency) by two trained English as a Second Language (ESL) teachers. For the purpose of this study we only considered the scores assigned by one rater since the other rater was unable to rate all the utterances and we intended to have as large a collection of rated utterances as possible for a good statistical analysis. As a result of this selection, there were 181 speech segments constituting 185 minutes of spontaneous speech samples. An attempt was made to rate the speech for fluency and phone

accuracy by two trained ESL instructors.

Table 4.1: TOEFL iBT proficiency score rubrics.

Score	Description
4	The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse.
3	The response addresses the task appropriately, but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas.
2	The response addresses the task, but the development of the topic is limited. It contained intelligible speech although problems with delivery and/or overall coherence occur and meaning may be obscured in places.
1	The response is limited in content and/or coherence is only minimally connected to the task or speech is largely unintelligible.
0	Speaker makes no attempt to respond or response is unrelated to the topic.

The score distribution is summarized in Table 4.2. In our experiments we use the data set in two ways: when studying measurements on the entire utterance we choose a set of rated utterances and call this set **Entire**, and when we use random 20 s snippets of the utterances, we call this set of snippets **Esnippet**.

Table 4.2: Distribution of human-rated fluency scores in the data.

	1.5	2	2.5	3	3.5	4
Number	5	37	67	45	22	5
%	2.7	20.3	36.8	24.7	12.08	2.7

4.2 Quantifiers of Fluency

Based on the results of previous studies and the fluency assessment rubrics for TOEFL iBT in [67], we identify two essential components of perceived fluency in a second language:

- the quantitative aspect of fluency, influenced by the speaker's ability to speak *effortlessly* and *quickly*;
- the qualitative aspect of fluency, influenced by the speaker's ability to communicate *effectively*—by getting his/her ideas across despite problems with the grammar, pronunciation and vocabulary.

Part of this study aims at understanding the extent to which human perceptions of fluency are influenced by the quantitative and qualitative aspects of speech. We anticipate that the understanding gained from this exercise will help us choose the right set of quantifiers of fluency for use in the automatic assessment system.

The aspects of fluency, as mentioned above, are ill-defined as measurement criteria. For instance, human raters have an implicit understanding of the notion of *effortless* and *quick* and rarely do they state that they based their rating on a certain speech rate of a certain syllables per minute. This is what renders the target quantity of fluency subjective. It is only by a suitable choice of quantifiers (here, via temporal variables and measures of lexical use) that we can attempt to make objective approximations of such subjective assessments. The quantifiers of oral fluency that we use in our experiments are dealt with next.

4.2.1 Measures of the Quantitative Aspect

Our goal is to quantify effort and speed of speech production by the speaker. Toward this end, we choose syllable-related information as well as information related to disfluencies obtained from the utterance, as relevant quantifiers. The quantifiers used here are the measures of temporal aspects of fluency that, when obtained automatically, have been shown to strongly correlate with native speakers' evaluation of second language speech [65, 19, 58, 67]. The measures that we use are also limited by the fact that the utterances

are of different durations. This calls for the use of time normalized measures to make the comparison fair. Denoting by d_1 the duration of the utterance without silent pauses and by d_2 the total duration of the utterance including silent pauses, we list below the quantifiers that we use:

1. Number of syllable nuclei (NUC),
2. Articulation Rate (AR) – number of syllable nuclei/d_1,
3. Rate of Speech (SR) – number of syllable nuclei/d_2,
4. Phonation/time ratio (PTR) – d_1/d_2,
5. Number of silent pauses per second (SPS) – number of silent pauses/d_2,
6. Total length of silent pauses (LOS),
7. Mean length of silent pauses (MLS),
8. Total number of silent pauses (SIL).

The only disfluencies that we consider here are the filled pauses. With this assumption, the quantifiers related to disfluencies are:

1. Number of filled pauses in the utterance (FP),
2. Number of filled pauses per second (FPS) – number of filled pauses/d_2.

The quantifiers are highly correlated among themselves but rather than looking for independent quantifiers, our intention was to seek a set that best correlates with the fluency scores while also having a good coverage of temporal aspects of speech production. In the experiments below, we measure all the quantifiers indicated above from the speech signal. However, owing to differences in the total duration of utterance in the speech segments, we decided to choose only those quantifiers (PTR, SR, AR, MLS, SPS, FPS) that are time-normalized for the automatic assessment module.

Central to obtaining the measurements of quantities listed above is the ability to:

- segment the speech signal into regions of speech and regions of silence: we use the intensity information obtained in the preprocessing stage to accomplish this;

- count the vocalic segments in the speech-portion of the utterance: we use voicing and intensity information to detect vocalic nuclei; and
- detect filled pauses: we used manually obtained filled-pause information in this experiment.

We now state the assumptions underlying the set of quantifiers chosen and then describe the algorithms of speech-silence segmentation and vocalic nucleus detection. Our assumptions are:

1. Repetitions and restarts are currently being considered as speech and the only disfluencies of interest are filled pauses.
2. Silent pauses are those segments of silence that are longer than 0.2 s in duration. These are unlike the utterance-internal pauses shorter than 0.2 s that occur as parts of word utterances.
3. Syllabic units are approximated by their vocalic nuclei. This approximation makes measuring articulation rate and speech rate easier since vocalic nuclei can be automatically detected with reasonable accuracy.

Our algorithm for speech-silence segmentation of the signal is outlined below:

- Use the average intensity as the threshold to classify each 10 ms segment of the signal as 0 or 1 depending on whether the intensity of the segment is lower or higher than the threshold.
- In the first pass of the segmentation make the silent pause decisions. Here segments of speech of duration less than 100 ms are spurious noise spikes and we regard them as parts of silent segments.
- In the second pass of the segmentation make speech decisions. Towards this, use the segmentation in the previous step and the convention that consecutive silent segments of duration less than 200 ms are parts of utterance-internal pauses to mark them as speech segments.

The input is the intensity information and the speech signal. The output is the signal segmented into speech and silence as well as count and duration information of silence and speech.

The vocalic nucleus detection algorithm is an open source Praat script [82] that uses a combination of intensity and voicing information to detect syllable nuclei. This counts the number of syllables in the speech segments available at this stage. While we exclude the filled-pause count from the syllables, the syllables corresponding to the hesitation phenomena are not excluded from the syllable count. Together, the information on silent pauses and the number of syllables yield the necessary quantifiers.

4.2.2 Measures of the Qualitative Aspect

As such, very few studies have focussed on this aspect of fluency as we noted in Chapter 2. Those that have sought to assess the effects of lexical use on a person's language proficiency have used certain measures of lexical richness (word-list free or word-list based). Here we would like to augment the understanding by studying the extent to which one's productive vocabulary influences fluency scores in spontaneous speech, and towards this we use some of the measures of lexical diversity already studied along with some that we propose based on studies in other domains. In choosing the set of quantifiers, we are limited by the short span of utterances (maximum length of an utterance is one minute). Another constraint that we impose on the choice of measures is that they be word-list-free measures, bearing in mind that one of the goals of this study is creating a system that is less reliant on language-specific resources. We now consider the set of quantifiers of lexical richness that we use in our experiments.

1. Number of word *tokens*, the total number of words used by the speaker (**TOK**);

2. Number of word *types*, the number of distinct words used (**TYP**);

3. Number of *hapax legomena*, the number of words used only once (**HAP**);

4. Guiraud index, the ratio of the number of types to the square-root of the number of tokens (**GI**);

5. Lexical density, the ratio of the number of content words to the total number of words used (**LD**);

6. Vocabulary growth rate, the ratio of the number of *hapax legomena* to the number of word tokens (**GR**).

Vocabulary growth rate given by

$$GR = \frac{HAP}{TOK},$$

is a new measure that we study in this context. Baayen in [37] defines this measure in the context of measuring the vocabulary growth rate from a sample of text and defines it to be the ratio of the expected number of words occurring once to the total number of words occurring in the text. We approximate the expected number of words occurring once by the observed number of words that occurred once. This is also the widely known estimate of probability mass of unseen words popularly known as the Good-Turing estimate of probability mass of unseen elements [52]. Intuitively, this measure gives the chance that the next word will be new. In other words, we construe this to be the chance that the speaker uses a new word.

Owing to the fact that the denominator is a quantity that is not constant across utterances with different number of words, this measure potentially has the same problems that the more commonly used measure, the type-token ratio, has. That is, the ratio is artificially affected by the number of tokens and is shown to be high in instances with low word token count. To counter this effect we make the following adjustment.

We first note the lowest number of word tokens uttered and call it the adjusted token count, which in our case is 30. Then we calculate the value of HAP for that token count. This is possible since we have a vocabulary growth curve (as also a growth curve of *hapax legomena*) as a function of word tokens. But rather than obtaining the value from the empirical growth curve, we resort to an interpolation method that assumes an underlying Binomially distributed sample. The computation involved is aided by the explicit computational modules available in ZipfR [83], a statistical package for R.

The measures are obtained from the available transcriptions of the utterances, where we consider a word token to be a string of letters delimited by space, without paying attention to the details of tokenization. For identifying content words, we use a list of function words such as one used in RANGE [61].

4.2.3 Correlation between Quantifiers and Fluency Scores

As seen in Section 4.2.1, a set of quantifiers of temporal aspects are obtained as measurements at the signal level. This requires obtaining silent pause information, syllable count information and filled-pause information. The segmentation process of dividing the signal into regions of speech and silence is very accurate with upwards of 99% accuracy. This renders accurate duration and count information on the silent pauses. The syllable detector performs well under noise-free conditions with accuracies over 90%. Thus the syllable count information is reliable as well. The filled pause information is manually obtained for the set of experiments. We thus have an accurate set of quantifiers of the quantitative aspects of fluency measured from the speech signal.

We first look at the means and standard deviations of the temporal measurements over the entire set of speakers to see how these measures vary with fluency scores. From Table 4.3 we see that the rate of speech for the speakers is 2.01 syllables per second, which is below the average of 4.3 syllables per second for conversational speech in native speakers of English [84]. This is in line with the fact the the speakers here are second language learners of English.

Table 4.3: Means and standard deviations of the quantifiers measured over complete utterances in the **Entire** data set.

	FP	**FPS**	**LOS**	**SPS**	**SIL**	**MLS**	**NUC**	**AR**	**SR**	**PTR**
mean	7.2	0.1	17.6	0.3	15.8	1.1	101.8	3.1	2.0	0.6
sd	5.9	0.1	8.6	0.1	5.3	0.4	37.5	0.3	0.4	0.1

From Table 4.4 we see that the mean values of the quantifiers show significant differences when compared across score classes. Here, by score class we mean the set of utterances getting a particular score. To see how this difference is carried over between fluency levels we consider classifying the utterances on the basis of their scores into two levels, *fluent* and *not fluent*, based on the mean score of 2.5. Thus, speakers with a score above 2.5 are considered fluent and those scoring 2.5 and below are considered not fluent.

Based on this categorization we analyze the measures over the utterances in both **Entire** and **Esnippet** to see if the two fluency groups differ significantly on these quantifiers. The mean values of the measures for **Entire** are shown

Table 4.4: Quantifier means are tabulated for each score class in the **Entire** data set.

Score	**FP**	**FPS**	**LOS**	**SPS**	**SIL**	**MLS**	**NUC**	**AR**	**SR**	**PTR**
1.5	4.6	0.20	16.39	0.37	16.40	0.98	79.60	2.80	1.78	0.64
2	8.9	0.171	20.31	0.33	17.16	1.21	95.7	3.01	1.85	0.61
2.5	8.3	0.170	16.35	0.32	15.40	1.09	97.7	3.04	1.99	0.65
3	6.3	0.121	17.53	0.31	15.55	1.11	103.3	3.1	2.08	0.66
3.5	3.9	0.071	17.05	0.30	15.95	1.09	122.3	3.23	2.20	0.68
4	5.4	0.108	11.48	0.28	13.4	0.91	121.6	3.32	2.53	0.76

in Table 4.5 and for **Esnippet** in Table 4.6.

Table 4.5: Quantifier means for the *fluent* set of utterances compared with the *not fluent* set for utterances in **Entire**. The differences in means are significant as evidenced by t-tests for each pair of means.

Level	**FPS**	**SPS**	**MLS**	**AR**	**SR**	**PTR**
Not fluent	0.26	0.34	0.87	3.02	2.14	0.70
Fluent	0.15	0.32	0.78	3.15	2.37	0.75
difference significant (at 5%)	yes	yes	yes	no	yes	yes

Table 4.6: Quantifier means for the *fluent* set of utterances compared with the *not fluent* set for utterances in **Esnippet**. The differences in means for the two fluency classes are significant as evidenced by t–tests for each pair of means.

Level	**FPS**	**SPS**	**MLS**	**AR**	**SR**	**PTR**
Not fluent	0.17	0.37	0.83	3.01	2.12	0.70
Fluent	0.13	0.35	0.74	3.18	2.37	0.74
difference significant (at 5%)	yes	no	yes	yes	yes	yes

T-tests for the significance of the difference between the means reveal that on the **Entire** data set, the differences between mean values of **FPS, MLS, SPS, SR** and **PTR** are significant. On the **Esnippet** data set, however, the differences in means of **FPS, MLS, AR, SR** and **PTR** over the two fluency classes (with the fluency classes defined as before) are significant (refer to Table 4.6). It is interesting to note that on an average the values of the quantities **PTR, SR, AR** and **MLS** for the 20 s snippets are the same as those for the entire utterance (compare Tables 4.5 and 4.6).

Box plots showing the distribution of the quantifiers for each score class of the data set are shown in Figures 4.1 and 4.2. This helps to note the

systematic variation of the quantities across score classes.

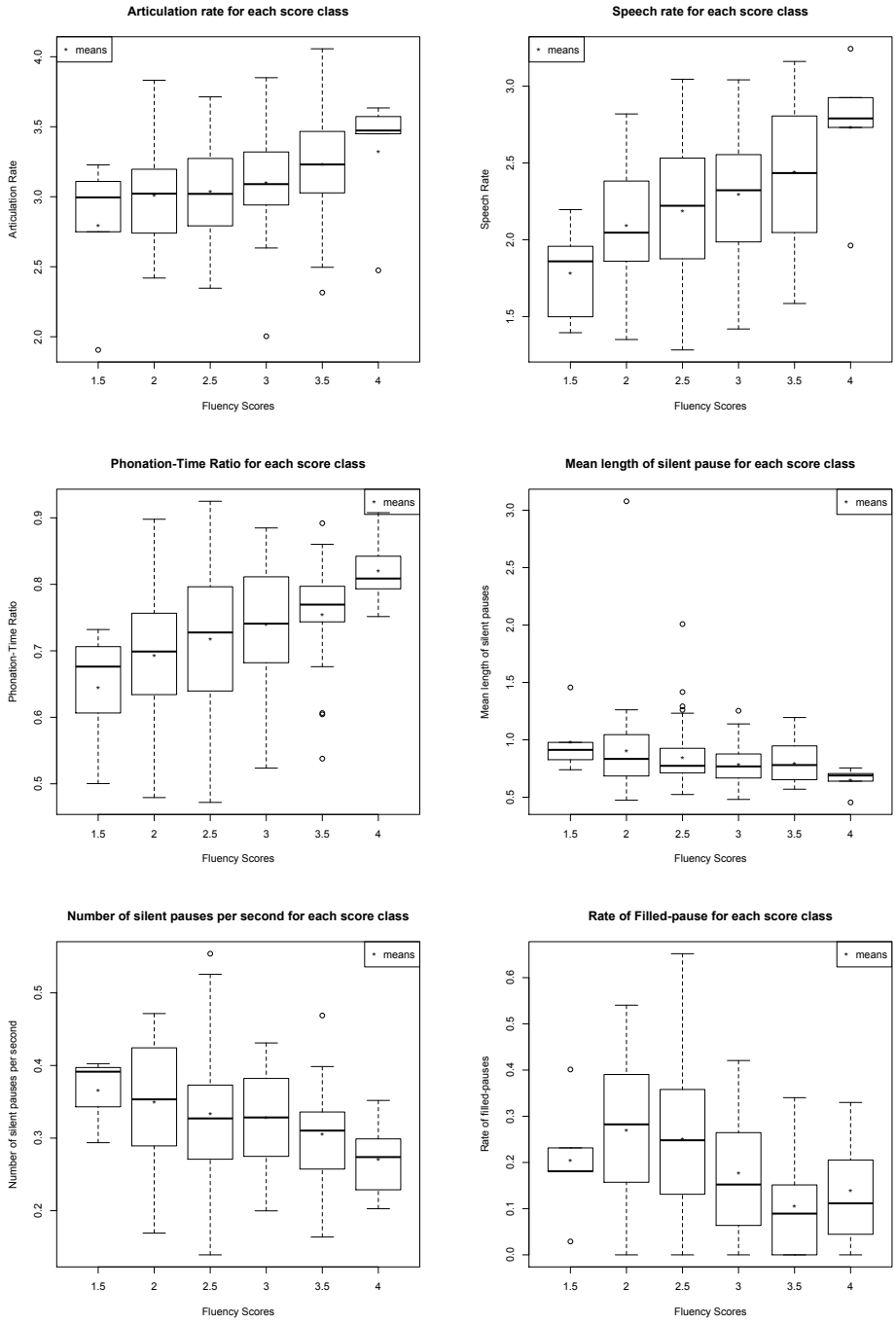

Figure 4.1: Plots of the different quantifier values for each score class along with the corresponding mean values. The quantifiers are obtained from the set of complete utterances **Entire**.

We measure the extent to which a quantifier predicts perceptions of fluency

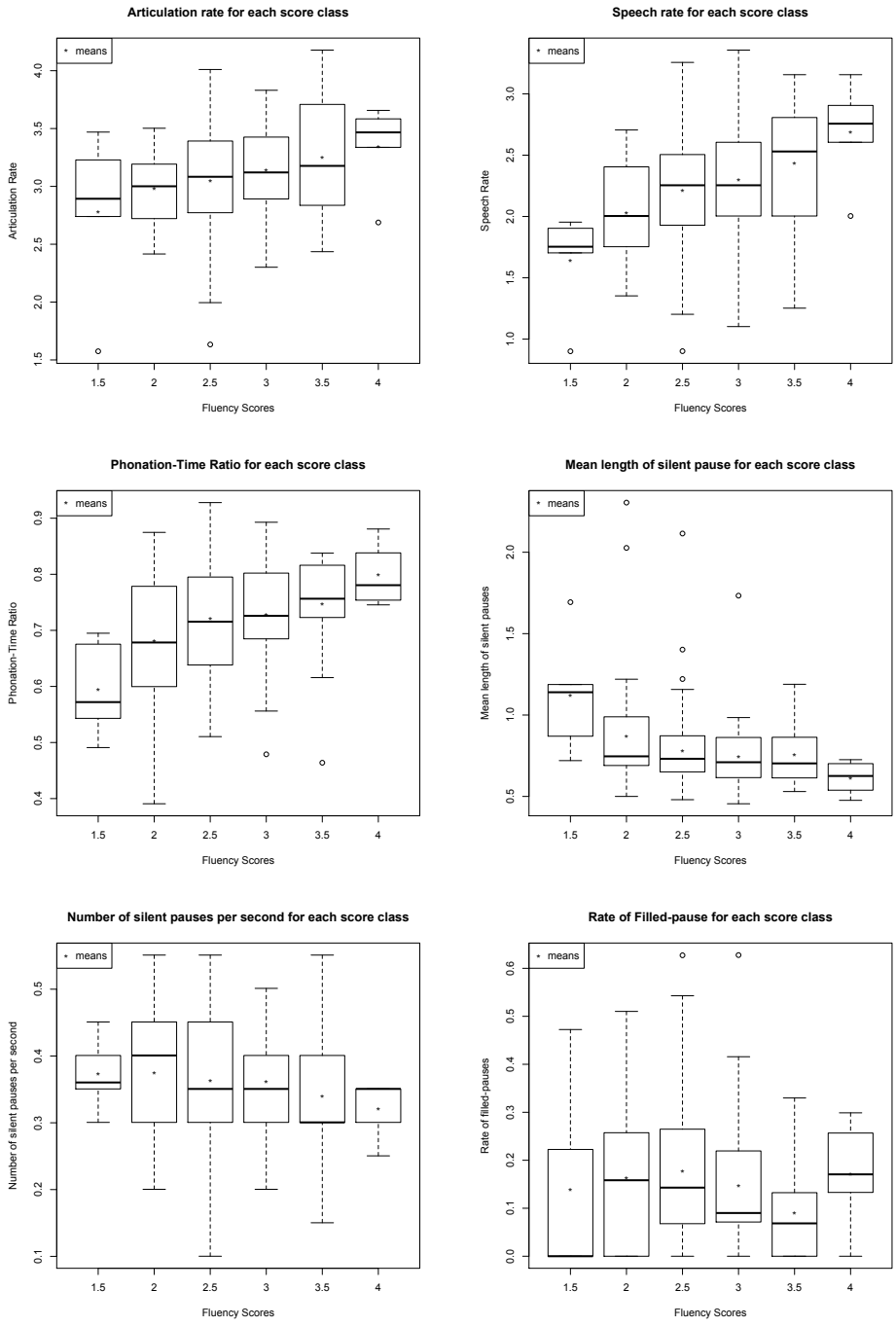

Figure 4.2: Plots of the different quantifier values for each score class along with the corresponding mean values. The quantifiers are obtained from the set of complete utterances **Esnippet**.

by the correlation coefficient of the quantifier with the fluency score. First we consider the quantifiers of temporal aspects of speech, following which we

consider the quantifiers of lexical use.

Quantifiers of temporal aspects of speech:

We consider the correlations between the quantifiers of temporal aspects of speech and the human-rated fluency score of the utterance in the data set **Entire**. In particular, with the scores on a 0-4 point scale (with 0.5 increments) we look at the Pearson's correlation coefficients of the means of the quantifiers at every score point with the scores.

The correlations among the features are listed in Table 4.7. We notice that

Table 4.7: Correlations between quantifiers in the **Entire** data set.

	PTR	**SR**	**AR**	**MLS**	**SPS**	**FPS**
PTR	1.00	0.82	0.24	-0.70	-0.68	0.18
SR		1.00	0.74	-0.57	-0.55	0.11
AR			1.00	-0.15	-0.15	0.04
MLS				1.00	0.004	-0.08
SPS					1.00	-0.19
FPS						1.0

high correlation (> 0.8) is seen between the pairs **PTR**-**SR**, while moderate correlation (> 0.5) is seen between **AR**-**SR**. We use the correlations between the quantifiers to guide us in their choice during the design of the automatic system as we will see in the next section.

We summarize the quantifier-score correlations in Table 4.8. From the table we notice that all the quantifiers show high correlations (positive or negative) with the human-rated scores. While **AR**, **SR**, and **PTR** are positively correlated with the scores, **SPS**, **FPS** and **MLS** are highly negatively correlated with the scores.

Quantifiers of lexical use:

The correlations of the quantifiers with fluency scores for the **Entire** data set are shown in Table 4.9 (we do not consider these measures for the **Esnippet** data set owing to the short duration of the utterances). We notice high correlations (> 0.8) between the quantities, TOK, TYP and HAP. This shows that the utterances that were perceived to be more fluent were more wordy, had larger number of word types and had more words that were used only once. We also notice that LD is moderately negatively correlated with fluency scores (although, with the current data set, the correlation is not significant). One possible explanation for this observation is that the more

Table 4.8: Correlations of the quantifiers with the human-rated fluency scores on the **Entire** data set. * indicates that the correlations are not significant at 5% level.

Quantifier considered	Correlation coeffient
FPS	*-0.75
SPS	-0.97
MLS	-0.97
PTR	0.98
AR	0.97
SR	0.98

fluent utterances are better phrased by the use of appropriate function words and are thus lexically dense. Experiments with more data would be needed to strengthen this observation.

Table 4.9: Correlations of the quantifiers of lexical richness with the human-rated fluency scores. * indicates that the correlations are not significant at 5% level.

Quantifier considered	Correlation coefficient
TOK	0.90
TYP	0.88
HAP	0.84
GR	*0.40
LD	*-0.69
GI	0.84

The table shows the correlation of vocabulary growth rate (GR) with fluency as being 0.4. Looking at the box plot of the distribution of the values of GR for each score class (refer to Figure 4.3), however, we can further qualify this correlation. We notice that the the score classes 2 and 3.5 seem to contribute towards the correlation and that in the other cases, the vocabulary growth rate seems to be negatively correlated with fluency scores. Thus, with the data set used we do not see good correlations between GR and fluency scores.

As a measure of lexical richness, the Guiraud index does not show the correlation that we expect to see with the fluency scores across all score classes. This is particularly so with the highest and lowest fluency scores,

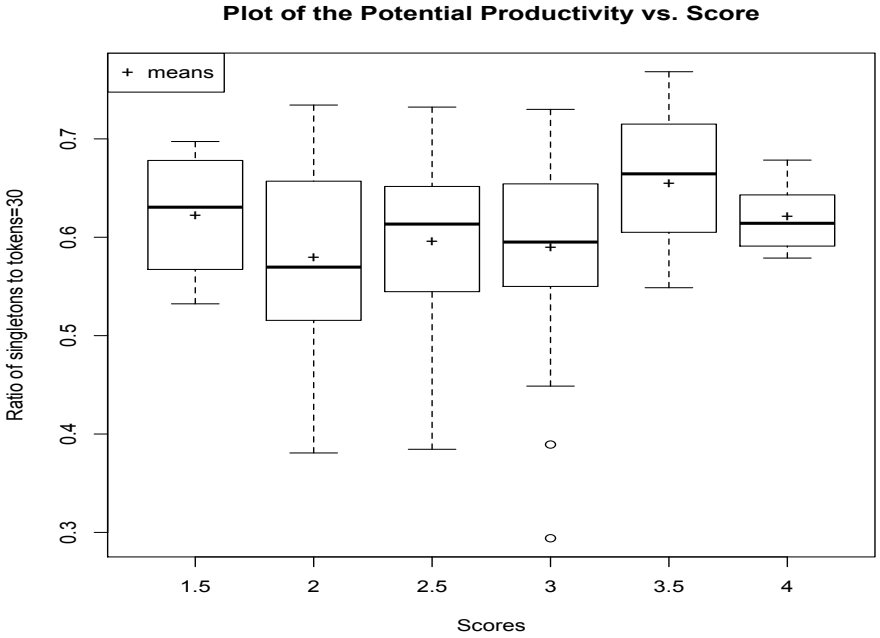

Figure 4.3: Box plots of values of vocabulary growth rate (GR) for each score class along with the corresponding mean values. The quantifiers are obtained from the set of complete utterances **Entire**.

where the number of word tokens is respectively very high and very low, as can be seen in Figure 4.4. Although we notice an overall correlation coefficient of 0.84, from the Figure 4.4 it is apparent that the contribution to the correlation is more from the central score classes than from the extreme ones. However, the Guiraud index appears to be well correlated (Pearson's correlation coefficient = 0.88) with the fluency scores in the score classes 2, 2.5, 3 and 3.5.

4.2.4 Conclusions on Quantifiers of Fluency

Based on the results mentioned above, we can draw the following conclusions:

- The quantitative aspects of fluency measured in terms of temporal quantifiers of speech are very good predictors of fluency, as has been observed in several previous studies. An added observation here is that the measures can be obtained by making direct signal-level measurements on the speech signal. While this is a step towards objective

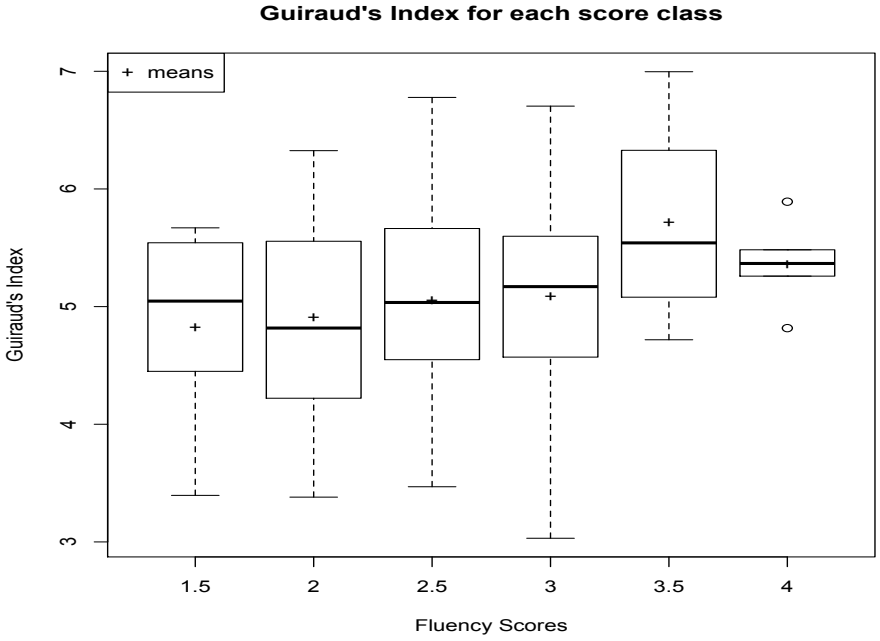

Figure 4.4: Box plots of values of Guiraud index for each score class. The quantifiers are obtained from the set of complete utterances **Entire**.

assessment of fluency, it is also a promising alternative to the resource-intensive ASR-based measurement.

- Lexical richness, as quantified by the number of word tokens and the number of word types, is a good predictor of fluency. However, the effect of other measures of lexical use on fluency as observed from the available data set seems inadequate to consider them as good predictors of fluency.

In the light of these conclusions as well as with our intention of studying the extent to which quantities derived from direct signal measurements are able to quantify fluency, we only consider the acoustico-temporal quantifiers of fluency as features in the automatic fluency assessment system to which we turn next.

4.3 Automatic Fluency Assessment System

In discussing the deficiencies of the state-of-the-art automatic fluency assessment system, we mentioned that one of the goals of the current study was to find alternative means of automatic fluency assessment. Ideally, an alternate system should be:

1. able to use classroom quality (some noise) or telephone quality (low bit rate) recordings;

2. able to handle recordings in a language independent manner (without relying on transcriptions) and hence be adaptable to data from a new language with minimum effort.

Such a versatile system will be suitable for a wide range of testing situations. For instance, such a system would be useful as a low-stakes language assessment module for use in a high-school classroom where possibly more than one second language is taught and tested.

We will now describe one such system, which, by using suitable estimators in a scoring model evaluates language fluency of a spoken utterance. A key component in the design of our system is the choice and measurement of suitable quantifiers of oral fluency that correlate well with the ratings of expert human raters. The set of quantitative variables is a set of direct signal-level measurements, acquired from the speech waveform. The measures are converted via logistic regression into an accurate estimate of human-rated fluency scores.

Thus, the practical advantages to our method are:

- Having signal level measurements as quantifiers affords a wide possibility of algorithms to measure the quantifiers.

- Without the need for transcriptions, our method can be used to analyze utterances in any language.

- Our automatic assessment module could be incorporated into a larger language proficiency testing system with very minor modifications.

We have seen in Section 4.2.1 the details of the manner of extracting low-level acoustico-temporal information from the speech signals. We now show

how we use the measurements to score fluency of spoken utterances automatically. Finally we evaluate the performance of our automatic system relative to the human ratings.

From the conclusions drawn in Section 4.2.4 we saw that temporal aspects of speech serve as good predictors of fluency. Based on this understanding we choose these quantifiers to represent aspects of speech production which will in turn serve as quantifiers of fluency in the design of the fluency assessment system.

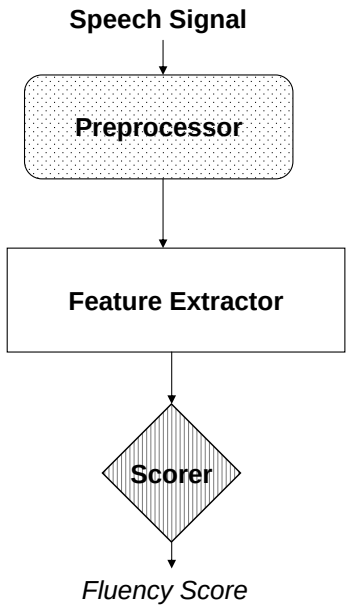

Figure 4.5: Architecture of the proposed automatic fluency assessment system.

4.3.1 System Architecture

In this section we describe the system architecture of the fluency assessment system where we understand the rationale and construction of each of its components, a sketch of which is shown in Figure 4.5.

The speech signal is first preprocessed, then relevant features are extracted

and sent to the scorer which then generates a fluency score for for the utterance. We will now consider the system components in detail.

Preprocessor The speech segment is in wav format. It is first resampled to 16 kHz and converted to mono. Then we obtain the pitch and intensity information for every 10 ms segment of the signal. The intensity information and the signal are then sent to the feature extraction in the next stage.

Feature Extraction The quantifiers used in our study are chosen so as to adequately represent aspects of speech production. In particular, the choice was based on

- the quantifier's relevance to the notion of speech quality;
- having a set of quantifiers with good coverage of the notion of speech quality as observed from previous related studies;
- empirical evidence from the data analysis phase that the chosen quantifier correlates well with the fluency score. In particular, we seek to quantify fluidity of expression and the associated listener's effort by using syllable-related information of the utterance as well as information related to disfluencies.

Scorer The scoring module acts as an estimator of the human scores given the signal-level measurements. It accepts the features generated in the previous module, generates the probability of the utterance being fluent given the features, and assigns a score to the segment as being fluent or not by thresholding on the probability.

We use a logistic regression model to generate the probability of fluency given the set of quantifiers as features. Advantages of the regression model are the simplicity with which the relation between the outcome and the features is represented and the interpretability of the resulting model in terms of the relative weights of the features. The feature coefficients of the model reflect the relative importance of the quantifiers governing the perception of fluency.

We denote by Y the binary random variable indicating whether an utterance is fluent or not, taking value 1 when the utterance is fluent and 0 otherwise. Let $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ denote the vector of the real-valued quantitative variables (maximum N=6 in our case), where each X_n is a variable such as **PTR** generated in the feature extraction module. The variables

are not statistically independent in our case and their underlying distributions are unknown but are assumed to come from a normal distribution. We also assume a linear relation between the predictors (here, the quantifiers) and the outcome (the fluency score) with no interaction between the predictors. Denote by p the quantity $P\{Y = 1|\mathbf{X} = \mathbf{x}\}$. Using the definition $\text{logit}(a) = \ln(\frac{a}{1-a})$ for $0 < a < 1$, the logistic regression model for approximating the probability of the utterance being fluent given the measurements is given by

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \ldots + \beta_N X_N, \tag{4.1}$$

where $\beta_i \in \mathcal{R}$ for $i = 0, \ldots, N$. The model is completely specified by the parameter vector $\beta = (\beta_0, \ldots, \beta_N)$. The parameters are obtained as those that maximize the likelihood of the observations in the data. We compute the parameters using the statistical software package R [85]. The output of the logistic regression model is the posterior probability of the outcome given the measurements. Using the posterior probability that the utterance is fluent we assign a score 1 (fluent) to the utterance if the probability is greater than 0.5 and a score of 0 otherwise. This way we convert the output posterior probability to a fluency score.

4.3.2 Performance Evaluation

We use 10-fold cross-validation to train and test the logistic regression model. The performance of the scoring module is judged in two ways:

- Accuracy: Since the outcome is considered a fluency score rather than a probability, the accuracy of the score in comparison with the human-rated scores (considered the target) is one performance criterion that we consider, defined as the percentage of the number of correctly assigned scores.

- Cohen's kappa measure: We use the κ-measure to assess the level of agreement between human-assigned and machine-assigned scores. Cohen's κ is given by

$$\frac{P(a) - P(e)}{1 - P(e)}, \tag{4.2}$$

  where $P(a)$ is the relative observed agreement and $P(e)$ is the hypothetical chance agreement between the two raters, calculated using the ob-

served scores. If the raters are in complete agreement then $kappa = 1$. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa \leq 0$.

We now shift our focus to discuss the performance of the scoring module. In this case, although the scores were on a 0–4 point scale, we converted them to a binary scale by thresholding on the mean score (here 2.5). This made the data suitable for a logistic regression scoring model.

We considered different configurations of the scoring model using different sets of quantifiers for the **Entire** data set. The best performance is seen with the measures **PTR**, **AR**, **MLS**, **SPS** and **FPS**. With this scorer, the automatic fluency assessment system based on the entire utterance emulates the human scoring procedure with an accuracy of 72.1%. In addition, a κ-score of 0.66 indicates good agreement between the human assigned and machine assigned scores.

Table 4.10 shows the coefficients and the relative importance of the different quantifiers used in the logistic regression scoring model. The relative importance of a variable is given by the change in the odds of the outcome per unit change in the variable, which for variable X_n is $odds\{Y = 1|X_n\} = \exp(X_n\beta_n)$, where β is the coefficient of X_n in the logistic regression fit. Here we do not consider interaction between the variables. We see that **AR** and **PTR** positively impact the score while **MLS**, **SPS** and **FPS** negatively impact fluency. Moreover, **AR** appears to impact the score to the greatest extent while **FPS** impacts the least.

Table 4.10: Coefficients of the quantifiers and their corresponding weights in the logistic regression approximation of human rated fluency scores based on **Entire**. The importance of a quantifier is given by the change in log odds in the outcome for a unit change in the quantifier value.

	PTR	**AR**	**MLS**	**SPS**	**FPS**
Coefficient	0.2070	1.5176	-1.9974	-6.4494	-7.7598
Importance	1.2300	4.5610	0.1356	0.0015	0.0004

4.4 Thin-Slice Assessment

Whether human-rated or automated, language proficiency assessment is typically done on spoken language segments several minutes in length. In this context, two natural questions arise: (1) Can we assess oral fluency by using portions of utterances? and (2) What measures obtained from a short slice of the utterance can be used towards automatic fluency assessment? In this section we look at our experiments aimed towards answering the above questions.

The experimental setup is the same as before, but the data set we use is the **Esnippet** set of random 20 s samples from the complete utterance. The proportion of the utterance represented by the snippet varies between 26.3% and 88.06% with a median of 44.6%. For this experiment, we disregard two snippets that are shorter than 20 s in length. The distribution of the proportions of the complete utterance represented by the random snippet is shown in Figure 4.6. We thus have a reasonable sample of snippets capturing

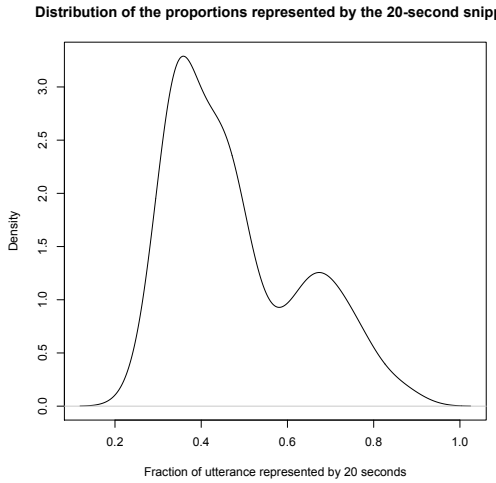

Figure 4.6: Distribution of the proportions represented by the 20 s random snippet as a fraction of the duration of the complete utterances.

not just a small portion of the utterance, but also a significant portion as well.

From Table 4.11 we notice that the correlations among the features is similar to those based on the **Entire** data set (refer to Table 4.7). As in the case of the **Entire** data set, we measure the extent to which a quantifier

Table 4.11: Correlations between quantifiers in the Esnippet data set.

	PTR	**SR**	**AR**	**MLS**	**SPS**	**FPS**
PTR	1.00	0.81	0.29	-0.67	-0.63	0.11
SR		1.00	0.79	-0.57	-0.48	-0.007
AR			1.00	-0.23	-0.11	-0.16
MLS				1.00	-0.11	-0.11
SPS					1.00	-0.19
FPS						1.00

Table 4.12: Correlations of the quantifiers with the human rated fluency scores. * indicates that the correlations are not significant at 5% level.

Quantifier considered	Correlation coefficient
FPS	-0.15
SPS	-0.93
MLS	*-0.91
PTR	0.95
AR	0.98
SR	0.98

predicts perceptions of fluency by the Pearson's correlation coefficient of the quantifier with the fluency score. Based on Table 4.12, we see that the quantity **FPS** is not seen to correlate well with the fluency score, but the other quantifiers are similarly correlated as in the **Entire** data set.

Noticing that the features aiding prediction of fluency scores with the **Entire** data set are also highly correlated with the fluency scores in the **Esnippet** data set, we use the same features in the scoring module to have a system that makes fluency assessment using snippets of the utterance as opposed to the entire utterance. Toward this end, we use the **Esnippet** data set to train the assessment system. The performance of the automated fluency assessment system for the different data sets considered in this experiment is tabulated in Table 4.13. We observe that the system based on **Esnippet** has an accuracy of 63.2%, whereas the system based on **Entire** has an accuracy of 72.1%.

In Section 4.2.3, we noted that on an average, the values of the quantities **PTR, SR, AR** and **MLS** for the 20 s snippets are similar to those for the entire utterance and that the quantifier-score correlations are similar in both

Table 4.13: Performance of the individual classifiers considered.

Data set	Accuracy(%)
Majority class (baseline)	60.3
Esnippet	63.2
Ewhole	72.1

cases. This suggests that the hypothesis that snippets could be as good as the entire utterance when the assessment is based on those quantities that show similar values is plausible.

In order to test whether the difference in performance of the two (complete-utterance based and snippet-based) systems is significant, we compare their accuracies using a version of McNemar's test described in [86]. We see that that the difference in performances is not statistically significant at the 1% level.

However, we notice that the majority class baseline in the data is 60.3% and looking at the 9% difference in accuracy results of the systems in this background suggests that there is a likelihood of the data being insufficient to draw conclusions on the difference in performance.

4.5 Discussion

The results of our study indicate that human perception of fluency is quantifiable by means of a set of variables. In particular, we saw that the measures that best predicted fluency were *articulation rate*, *phonation-time ratio*, *number of silent pauses per second*, *mean length of silent pauses*, and *number of filled pauses per second*. This result is not new since the use of these quantifiers as predictors of fluency has been well documented in the literature [19, 55, 56, 65, 87, 88, 89, 90]. We summarize in Table 4.14 our results alongside those from previous studies in an attempt to compare the best predictors of fluency (in terms of having the strongest correlation with the fluency scores) from available studies.

From Table 4.14 we see that our results are comparable with those previously concluded, in that rate of speech is seen as the best predictor of fluency. We also have phonation time ratio as having a strong correlation with fluency scores, which is similar to the results observed by Cucchiarini

Table 4.14: Comparison of the most correlated quantifier of fluency in available studies alongside our results.

Study cited	Quantifier (best predictor)	Correlation coefficient
Current study	PTR and SR	0.98
Yoon [67]	SR	0.58
Mizera [58]	SR	0.80
Cucchiarini et al. [65]	SR	0.97

et al. [65]. Their study concluded that phonation-time ratio was the second best predictor of fluency for spontaneous speech. The correlations of these quantifiers with fluency scores for the study by Zechner et al. [19] were not mentioned and are hence not available for comparison.

Another interesting observation was that measures of disfluency showed lower correlation with fluency scores as compared to the other measures. In addition they served as poor predictors in the presence of the well correlated measures. This could be interpreted to mean that perceptions of fluency are primarily governed by measures of temporal aspects of speech that reflect speech production and are only secondarily affected by disfluencies.

The novelty of our results lies in that the set of measures is obtained automatically by direct signal-level measurements. Automatic measurements of the quantifiers have been carried out in [19, 65] by the use of ASR which, although is state of the art, has accuracies that are far from being acceptable. Direct signal-level measurements of the quantifiers that we use render even telephone quality speech usable for automatic assessment. This makes our approach usable in many more second language learning and testing environments than are currently possible.

With few studies available on assessing the effects of lexical use on fluency scores, a comparison such as the one above for temporal features of speech could not be done. However, based on studies assessing the effect of lexical richness on overall language proficiency using word-list free measures of lexical richness we can make the following comparisons.

- While exploring measures of vocabulary richness in semi-spontaneous French speech, Tidball and Treffers-Daller [63] observe a correlation of 0.75 between the Guiraud index and scores of language proficiency. Although we observe a correlation of 0.84 with our data, owing to the

paucity of observations in the lowest and highest score classes as well as the effect of the word tokens, it is hard for us to make conclusive statements on this measure.

- The measure of vocabulary growth rate (GR) from a short utterance seems to be inappropriate for spontaneous speech of lengths considered in this study. It is possible that this measure quantifies lexical richness better when the number of word tokens is much larger than what we have considered here. It is up to further experiments to verify this.

- As measures of vocabulary use, counts of word tokens and word types were seen to increase with increased proficiency levels in [91]. This has been interpreted to mean that increased proficiency levels are associated with speakers using more words (for increase in number of tokens) and words in a wider range (for increase in number of word types). Taking fluency scores as measures of language proficiency, our results are in line with these observations, since we see that the correlation of TOK with fluency scores is 0.90 and that of TYP with fluency scores is 0.88.

- Lexical density (LD) is seen to be negatively correlated with fluency scores (the correlation, however, is not significant at the 5% level). Although this can be interpreted to mean that more fluent utterances tend to be more structured and hence tend to use more function words when compared with less fluent utterances, more experimentation is necessary to make conclusive statements on this. As pointed out by Laufer and Nation [61], this measure being influenced by the number of function words relative to the total number of words used is perhaps more indicative of the structural characteristics of an utterance than of the lexical richness.

We then explored the design of an automatic fluency assessment system that, by way of extracting quantifiers of fluency as acoustic measurements of the speech signal, uses the measurements in a logistic regression framework and generates a fluency score that shows reasonable agreement with human-assigned scores. The resulting system is shown to be a prototype of an alternative to the ASR-based automatic fluency scoring module studied by Zechner et al. in [19]. The information on the performance of the fluency

module alone of the *Speechrater* not being available (this being the only available automatic fluency assessment system from previous work), we are not in a position to assess the performance of our system in comparison with other systems. However, judging by the good agreement on the human-machine scores ($\kappa = 0.66$) of our system, we have reason to believe in the potential utility of our system in low-stakes testing scenarios such as classroom settings as well in scenarios where language-specific resources are scarce.

The observation that the values of the quantities **PTR, SR, AR** and **MLS** for the 20 s snippets are similar to those for the entire utterance suggests the possibility that factors affecting perception of fluency are perhaps not results of a global phenomenon, but possibly somewhat local. It will be interesting to see the different utterance durations where this observation holds.

Finally, our results showing that the difference in performance (accuracy) between assessment based on the entire utterance and that made using thin-slice segments is apparently substantial but is not statistically significant at the 1% level, suggests further experimentation with more data. This problem is interesting not only in automatic language testing domains, but also has potential implications in other domains where attributes of temporal effects of speech would be quantities of interest.

CHAPTER 5

VARIABLE SELECTION MODEL FOR ADVERTISEMENT PREDICTION

We consider the problem of predicting the probability of a click for an advertisement when the outcome of a click or no-click is expressed by means of a set of variables. The 42 variables represent measurements such as:

- query related quantities such as `MatchedKeyword` which stands for the number of words in the query that match with the keywords associated with the ad;
- advertisement related quantities such as `ListingID`;
- user related quantities such as `IP address` and `Age`; and
- and some general quantities such as `DayOfWeek`.

The *outcome* associated with this set of measurements is a one or a zero indicating whether the advertisement was clicked or not. The data was obtained from Microsoft's proprietary query logs over a period of several months.

The problem of estimating the probability of a click given a set of variables can be viewed as one of estimating the conditional expectation of the outcome given the values taken by the associated set of variables. If the variables were very few, then we could empirically learn the *joint* statistics between them and the single outcome. This could then be used to design an algorithm that predicts the outcome given the variables. Here, however, the number of variables is very large. Thus there are issues of data sparsity in the observed data. Some variables take values in a large set of choices, many of which may not have been observed in the sample. Even those that have been observed may only have occurred a small number of times. This makes the task of finding a good estimate of the joint statistics from the training data hard.

However, we have empirical estimates of the joint statistics between the outcome and each of the variables from the observed data. Our approach is to

extrapolate these estimates to obtain the necessary conditional expectation of the outcome given the variables.

The problem of modeling ad click-through rates by generating a set of features and then using them in a logistic regression model has been studied by [69]. In the current work we first perform variable selection and then use the variables thus obtained to fit a logistic regression model to predict the probability of a click on an ad.

Several approaches to variable selection have been proposed which can be broadly classified as those that perform variable ranking and those that perform forward or backward selection of nested subsets [22]. In this paper, we take the forward selection approach to perform the task of variable selection sequentially. We use a logistic regression model for extrapolating the estimates of the joint statistics between the outcome and each of the variables thus obtained to that of the conditional expectation between the outcome and ensemble of the variables. We compare the performance of two models: one constructed using the set of variables obtained by the incremental search heuristic and another using the set of variables obtained by an exhaustive search. We show that the proposed model shows near optimum performance with minimum computational effort.

5.1 Modeling the Probability of a Click

We now introduce some notation used in our model. We denote by Y the binary random variable indicating a click or no-click on an advertisement, taking value 1 when the advertisement was clicked and 0 otherwise. Let

$$\mathbf{V} \stackrel{\text{def}}{=} (V_1, V_2, \ldots, V_N)$$

denote the vector of variables (N=42), where each V_n is a variable such as **Listing ID**. The variables are not statistically independent and their underlying distributions are unknown. However, from the set of observations we obtain empirical estimates of the marginal distribution of the outcome given each of the variables which is $\Pr(Y = 1|V_n = v_n)$. To account for unseen values of the variables we perform adequate smoothing (e.g. add-constant).

Our task is to find:

1. a good model that represents the expected value of the outcome given the values taken by the associated variables;

2. a subset $(V_1, \ldots, V_K)$ of variables $(K < N)$ such that the variables incorporated in the model above best represent the outcome.

Each example (all assumed to be independent and identically distributed) in the available data was the vector of values taken by the variables and the outcome, a 0 or a 1. The available data is split into a training set with M=80,000 examples, a validation set with 10,000 examples and a test set with 10,000 examples.

5.1.1 The Statistical Model

Predicting the outcome given a set of variables is the problem of estimating $\mathbb{E}(Y|\mathbf{V} = \mathbf{v})$. Since Y is a binary variable this is equivalent to estimating $P\{Y = 1|\mathbf{V} = \mathbf{v}\}$. However, we have estimates of the marginal distribution of the outcome given each of the variables $P(Y = 1|V_n = v_n)$, which we would like to extrapolate to obtain the required estimate $P\{Y = 1|\mathbf{V} = \mathbf{v}\}$. The target function of the set of variables which we estimate is hence a probability distribution over the set of variables rather than a single label (1 or 0). We do this by adopting a first order logistic regression model to express the approximation.

Denote by p the quantity $P\{Y = 1|\mathbf{V} = \mathbf{v}\}$. Using the definition, $\text{logit}(a) = \ln(\frac{a}{1-a})$ for $0 < a < 1$, we denote by X_n the quantity $\text{logit}(P(Y = 1|V_n = v_n))$. Then the logistic regression model for the required approximation is

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \ldots + \beta_N X_N, \tag{5.1}$$

where $\beta_i \in \mathcal{R}$ for $i = 0, \ldots, N$. The model is completely specified by the parameter vector $\beta = (\beta_0, \ldots, \beta_N)$. The parameters are obtained as those that maximize the likelihood of the observations in the data. We compute the parameters using the statistical software package *R*.

5.1.2 The Performance Criterion

The problem of predicting the click-through rate being one of estimating a probability distribution, rather than one of classification, our performance criterion is optimizing an objective function that measures the goodness-of-fit of the model. Accordingly, we seek a probability distribution that is a good fit to the observation. We thus choose the model that maximizes the log-likelihood of observations.

We have described the steps in modeling the prediction of the probability of a click given the values taken by the associated variables. Our next focus will be on methods of choosing the set of variables that go into the model.

5.1.3 Variable Selection

The task of variable selection deals with choosing a subset of variables that together have good predictive power. Our approach can be seen as a wrapper method [92] where we use the learning method itself to score subsets of variables according to their predictive power using examples in the training data. The learning method uses the validation data to learn the parameters associated with the best subset of variables and presents the hypothesis. The resulting hypothesis is then evaluated on a test set. This approach requires one to define:

1. a learning method;
2. a way of searching the space of all possible subsets;
3. a way to guide the search procedure by using the performance of the learning method.

In our work, we choose logistic regression as the learning method. The possible subsets of variables are scored by using them in a logistic regression fit to the training data and assessing their performance in terms of the log-likelihood of the unseen data using the resulting model. We choose an exhaustive search and a hill-climbing heuristic as two ways of searching the space of all possible subsets of variables. We stop the search procedure when the increase in log-likelihood by the addition of one more variable is not statistically significant.

Having considered the logistic regression model and the performance criterion in the previous section, we now describe the ways in which we search the space of all subsets of variables in detail.

5.1.4 Exhaustive Search

We first perform an exhaustive search of the best subset of k variables. Toward this end, we initially obtain subsets of variables of size k for $k = 1, 2, \ldots, 42$. We then obtain logistic regression models for each of the subsets. Since our performance criterion is maximizing the log-likelihood, for each k we choose the best model as one having the maximum log-likelihood among other models with k terms. This set of k terms is then chosen as the best subset of variables of size k.

The computational cost incurred in the process of such an exhaustive search is shown in the Table 5.1.

Table 5.1: Computational cost of exhaustive search for the best k-subset from 42 variables

k	*Subset size*: $\binom{42}{k}$	*time*
1	42	24 seconds
2	861	7.1 minutes
3	11480	1.6 hours
4	111930	15.5 hours
5	850668	5 days
6	5,245,786	30 days
7	26,978,328	156 days
8	118,030,185	683 days

Owing to the high computational cost of the exhaustive search we limited our search up to $k = 5$.

As can be inferred from the table, an exhaustive search through the model space in the manner described above becomes infeasible very quickly when the number of variables is large. We thus resort to a hill-climbing search heuristic that provides a better path through the search.

5.1.5 Incremental Search

The incremental method that we use is a method of forward selection that employs a greedy search strategy. Here variables are progressively incorporated into larger sets yielding nested subsets of variables. The search begins with an empty set of variables. It then adds a variable to the set in such a way that the resulting logistic regression model is one that most improves the fit (in our case, has the maximum log-likelihood). At every step of the search process, since all the models obtained by adding one more term to the existing model have the same number of terms, we are, in turn, seeking to compare models by their individual ability to best explain the training data. We stop the incremental process when the increase in log-likelihood by the addition of one more variable is not statistically significant.

To test whether the addition of a variable to an existing model significantly improves the fit, we consider testing the corresponding likelihood ratios. Let M_k denote the model under test and M_{k+1} the extended model containing an additional variable. Then the quantity

$$D = LL_{k+1} - LL_k \tag{5.2}$$

is distributed like χ^2 with 1 degree of freedom when the likelihoods are computed for a large number of observations that are distributed independently according to a binomial distribution [93]. We then test the hypothesis that D = 0 versus the hypothesis that D ¿ 0.

The algorithm is as follows:

1. `Begin with a logistic regression model with only` β_0. `Set` $k = 0$.

2. `Set` $k = 1$

   - `Add one variable to the existing` $k-1$`-term model such that the resulting` k`-term model has largest LL.`
   - `Set` $k = k + 1$
   - `If LL(k+1) - LL(k) is not statistically significant stop.`

The resulting K-term model is then evaluated on the *test* data. We perform an incremental search through the model space using the algorithm for $k = 1, \ldots, 8$.

5.2 Evaluation

We consider the evaluation of two aspects of our experiment.

1. the performance of the proposed method of variable selection;
2. the quality of the logistic regression model using the best set of variables as a model for predicting ad click-through rates.

5.2.1 Evaluation of the Incremental Search

Figure 5.1 shows how the models obtained by the proposed incremental search compare with those obtained by the exhaustive search when evaluated on the validation data.

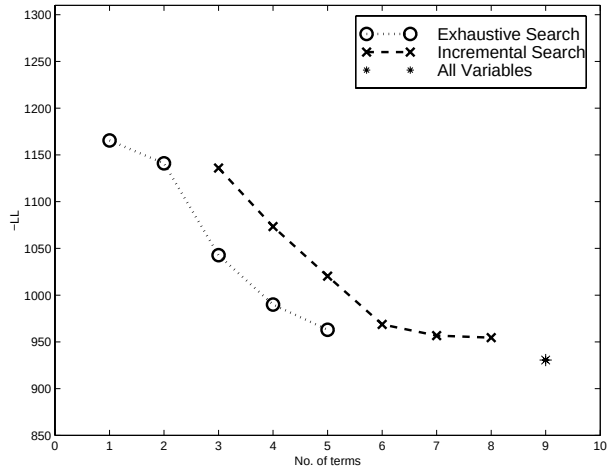

Figure 5.1: Plot of -LL values on the validation set of the models with subsets of variables obtained using exhaustive search and the incremental search.

Even without the intermediate models from the exhaustive search, we see that the incremental model with 7 variables is beginning to show its near optimal predictive power by comparing its log-likelihood with that of the model with all the 42 terms. The best subset of variables is obtained using the training data, and the associated parameters of the corresponding logistic regression model are obtained using the validation data. Finally, the models are compared on the test set and the results are plotted in Figure 5.2. We see that the 7-term model obtained using the incremental search performs as well as the best 5-term model.

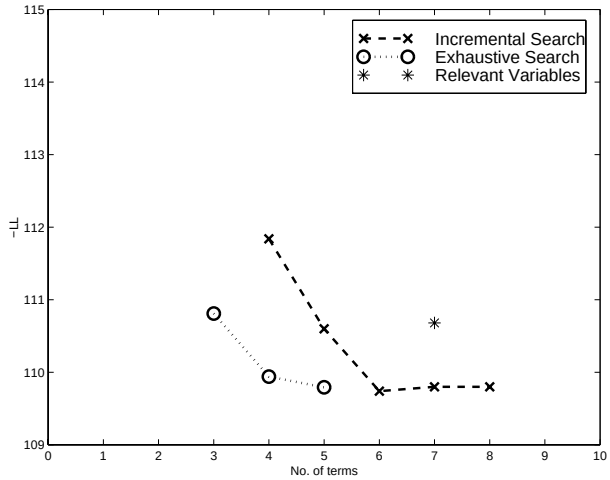

Figure 5.2: Plot of -LL values on the test set of the models with subsets of variables obtained using exhaustive search, the incremental search and statistical significance

The computational gain (as compared with the exhaustive search) by the use of the forward selection procedure can be seen from the Table 5.2.

Table 5.2: Computational cost of incremental search for the best k-subset from 42 variables. In contrast with that of exhaustive search, the computation with forward selection does not become prohibitive with k.

k	*subset size*	*time*
1	42	24 s
2	41	20.5 s
3	40	20 s
4	39	19.5 s
5	38	19 s
6	37	18.5 s
7	36	18 s
8	35	17.5 s

5.2.2 Evaluation of the Prediction Model

As a point of comparison with the chosen set of variables that serve as good predictors, we were interested in listing the variables that were relevant in fitting a logistic regression model of the probability of the outcome. With this in mind we look at the statistical significance of the individual parameters in the model with all the 42 variables. A high level of significance of a

parameter indicates that the corresponding variable is highly relevant to the model fit. This resulted in 7 variables being relevant. The resulting set of 7 relevant variables is listed here, classified into four groups based on domain knowledge:

1. Query related variable: **CleanQuery**;
2. Advertisement Display Position related variables: **ML.1, ML.2, ML.3** (ML = MainLine);
3. Advertisement related variable: **ListingID**;
4. User related variable: **UserID, IP4**;

However, despite their significance, the variables do not together form a set of good predictors as we can see from the figure. What we see as being a good set of 7 predictors (obtained from the incremental search procedure) includes some of the relevant variables—**ListingID**, **CleanQuery**, **UserID** and **IP4**—as well as some of the "non-relevant" variables such as **Subcategory**, **DayOfWeek**, **Ppzip**. This could be attributed to the joint effect of the less relevant variables in conjunction with the more relevant ones. Another likely explanation is that some of the relevant variables and the less relevant ones are correlated.

We now consider how the logistic regression model for predicting ad click-through rate can be interpreted as explaining the underlying process. The coefficients of the model render themselves to easier interpretation owing to the fact that the dependent variables in the regression model are logit transformations of the conditional probabilities of the outcome given the corresponding measurement of the variable. In Table 5.3 we refer to the model with 7 terms and the coefficients of the associated variables. Some of the variables still fall into the categories of query related, advertisement related and user related. From the set of variables with positive coefficients we see that `ListingID`, `CleanQuery`, `UserID` and `IP4` positively influence a click, while `DayOfWeek`, `Subcategory` and `Ppzip` negatively influence a click. The negative coefficients can be attributed to the correlation between the variables; not only do the variables individually influence the outcome, they jointly affect the outcome as well. This suggests a lack of conditional independence given the outcome among the seemingly unrelated variables.

Table 5.3: Models for variable subsets obtained by the incremental method $inc4, inc5, inc6, inc7$ with the model *rel* obtained using just the relevant variables.

Variable	**inc4**	**inc5**	**inc6**	**inc7**	**rel**
intercept	0.9	0.1	0.1	0.1	4.7
ListingID	0.7	0.8	0.8	0.8	0.7
CleanQuery	0.4	0.3	0.3	0.3	0.3
Subcategory	-0.8	-0.1	-0.1	-0.1	-
UserID	-	0.9	0.8	0.8	0.7
DayOfWeek	-	-1.0	-1.2	-0.8	-
IP4	-	-	0.4	0.4	0.2
Ppzip	-	-	-	-0.4	-
ML-1	-	-	-	-	-2.02
ML-2	-	-	-	-	-1.46
ML-3	-	-	-	-	-1.33
-LL	111.84	110.60	109.74	109.80	110.68

Table 5.4: The models for variable subsets obtained by the exhaustive search—the coefficients of the variables, and the -LL for the models.

Variable	**best-3**	**best-4**	**best-5**
intercept	-0.8	-0.8	0.0
ListingID	0.9	0.9	0.8
UserID	1.0	0.8	0.8
DayOfWeek	-1.2	-1.2	-1.3
IP4	-	0.4	0.4
CleanQuery	-	-	0.2
-LL	110.81	109.94	109.79

We also note that the models have approximately the same coefficients for the shared variables. Additionally we observe that this property continues to hold in the models obtained via exhaustive search. In terms of log-likelihood we also see that the search heuristic works almost as well as the exhaustive search (refer to Table 5.4).

We now summarize observations on the performance of the logistic regression model with seven variables obtained using our proposed search heuristic (cf. Figure 5.2):

1. Its performance compares favorably with that obtained by an exhaustive search.

2. Its performance is better than that obtained with only relevant terms.

CHAPTER 6

CONCLUSIONS AND FUTURE DIRECTIONS

In this chapter I will summarize the results from each of the chapters and arrive at conclusions based on the discussions in the respective chapters. Although there is an underlying theme binding the individual chapters as explained in Chapter 2, the conclusions that we draw in this study will pertain to the problems studied and appear as disparate owing to the nature of the experiments conducted.

6.1 Vocabulary Size Estimation

We showed mathematically that the Good-Turing vocabulary size estimator with the assumption of equally likely rare events always underestimates the vocabulary size of the population when rare events are not equally rare. From empirical observations with natural language corpora, we saw that the degree of underestimation can be as high as 30%. This is a serious shortcoming of an estimator that is deemed widely applicable.

To remedy this shortcoming, we propose an estimator of vocabulary size. Our proposed estimator makes a key modeling assumption about the distribution of large number of rare events. We make use of an earlier result that the scaled frequencies converge to continuous mixtures of Poisson distributions. This allows us to model the number of unseen events as a linear combination of the frequencies, making it a nonparametric estimator. We then showed that this estimator is statistically consistent. Its usefulness for estimation involving natural language corpora is seen in the results of our simulation where we observe that its performance is comparable to and at times even surpasses that of state-of-the-art estimators.

6.2 Automatic Fluency Assessment

Based on the experiments and results of the present study we draw the following conclusions.

Measures of lexical use quantified by the number of word tokens, number of word types and the number of *hapax legomena* are good predictors of fluency scores. The observed effects of lexical density, Guiraud index and vocabulary growth rate on fluency scores were insufficient to draw conclusions about their influence on perceptions of fluency scores. This we attribute to the skewed distribution of our current data set with very few lowest and highest score points.

Measurements of objective properties of speech, indicative of temporal aspects of speech production obtained directly from the speech signal using low-level measurements, serve as good quantifiers of human-perceived fluency. In particular we found that articulation rate, phonation-time ratio, mean length of silent pauses and the number of silent pauses per second are well correlated with fluency scores. Combining these measures in a logistic regression framework, we experimented with a scoring model that predicts the human scores as accurately as possible. The resulting system showed that the measure that contributed the most to the automatic scoring was *articulation rate*, followed by *phonation-time ratio*. In comparison to these measures, those of disfluency not only were less correlated with fluency scores but also served as poor predictors in the presence of the well correlated measures. This suggests that quantities that measure the wordiness of speech are better predictors of fluency than are quantities measuring disfluencies.

The low-level signal measurements make objective methods of fluency assessment using the quantifiers less reliant on ASR. This renders automated systems based on such objective methods usable in testing scenarios where the language being tested does not have enough resources for building an accurate ASR system.

We noticed that aspects of fluency that are measurable from the speech signal are not just global measures but are also measurable locally, provided a long enough snippet is chosen. More studies are need to establish the limit on the duration of speech needed for making such inferences.

Motivated by studies in social psychology, our experiments on assessing the judgment accuracy based on thin-slices of the spontaneous utterance showed

that the data was insufficient to draw conclusions about the performance difference between thin-slice based automatic assessment and that based on complete utterances.

6.3 Variable Selection for Click-Through Rate Prediction

Effective web advertising relies on a good estimate of an advertisement's click-through rate, as a function of several variables. We have shown the utility of a subset selection procedure that efficiently selects the key subset of variables that are most relevant to the outcome (clicking of the advertisement). We use this algorithm in the context of a logistic regression model that provides an estimate of the click-through rate. Numerical results demonstrate the efficacy of our approach, even when compared to a brute force exhaustive search for variable subset selection. We believe that the computationally efficient search heuristic of selecting a subset (that best represents the underlying relation between the variables and the outcome) of variables from a large set, is quite general and can be applied in other domains as well.

6.4 Contributions

The intricate challenges involved in the estimation problems in speech and natural language call for broad approaches to solving them, ranging from domain specific experimental work to the design of provably optimal algorithms. Each of the three efforts described in this thesis falls in a different interval of the solution range. Specifically, we summarize below our contributions in each of the three studies:

- The problem of vocabulary size estimation based on observations is a classical one (with at least 60 years of track record in the literature). Our main contribution is to take a renewed look at this problem by focusing on an important property of natural language corpora. Our main result is a nonparametric estimator of the underlying vocabulary size. We study its use in an explicit application—that of estimating the underlying vocabulary size of large natural language corpora—where

we show that it is not only computationally simple but also compares favorably with the state-of-the-art estimators in terms of performance [94].

- In designing a system for automatic fluency estimation, we showed that human perceptions of oral fluency are primarily influenced by temporal aspects of speech and only marginally influenced by aspects related to vocabulary richness and lexical use. We also showed a potential automatic fluency assessment system that is not reliant on ASR. In this interdisciplinary effort, joint discussions with linguists, psychologists and human judges of fluency themselves were very useful. As such, the output in this effort is domain specific.

- Finally, the problem of estimating the click-through-rate of advertisements was suggested in an industrial research setting (Microsoft Research), where it was a very topical problem [95]. Our approach to solving the problem allowed for a computationally simple implementation that outperformed state-of-the-art models of click-through rate prediction.

6.5 Future Directions

Although the problems considered in this thesis have been thoroughly studied, there remain several avenues that were unexplored owing to time constraints. Below I highlight possible avenues for future exploration in the problems considered.

Vocabulary Size estimation:

- In vocabulary size estimation for natural language corpus we only considered a sample as a bag of words with no underlying structure. The LNRE model we considered paid no particular attention to the generative process that governs the individual word and sentence formation. Future studies could be proposed where the effect of the underlying grammar and hence the constraint of a linguistic structure is taken into account while modeling rare events.

- In characterizing rare events, we considered a low probability event to

be one whose probability was bounded by constants. While we did not qualify this constant any further, one could explore the nature of this constant as a function of the underlying process. For instance, considering the process of generating typographical errors as a process that generates rare events, one could study the nature of the constants bounding the probabilities of occurrence.

Automatic fluency assessment:

- For our study, we relied on signal-level measurements as quantifiers of fluency, which are by no means exhaustive. Future studies could improve upon this by enlarging the set of quantifiers that affect perceived fluency.

- Owing to limited available data, this study considered fluency scores without regard to differences in proficiency levels of the speakers. Further studies could explore how different fluency levels over various degrees of language proficiency can be established using the measurements that we propose.

- Studying assessment based on thin-slices, one could investigate the limits of thin-slice duration that can provide measurements adequate for fluency assessment. This would mean experimenting with various snippet lengths for automatic assessment.

- Detection of filled pauses by signal level measurements, is another aspect that could be studied considering the relevance of the problem of detecting filled pauses in several speech processing applications.

REFERENCES

[1] B. Efron and R. Thisted, "Estimating the number of unseen species: How many words did shakespeare know?" *Biometrika*, vol. 63(3), pp. 435–437, 1976.

[2] R. A. Fisher, A. Corbet, and C. B. Williams, "The relation between the number of species and the number of individuals in a random sample of an animal population," *Journal of Animal Ecology*, vol. 12(1), pp. 42–58, May 1943.

[3] A. Stam, "Statistical problem in ancient numismatics," *Statistica Neerlandica*, vol. 41, pp. 151–173, 1987.

[4] M. Charikar, S. Chaudhuri, R. Motwani, and V. Narasayya, "Towards estimation error guarantees for distinct values," in *Proc. of the Nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, Dallas, TX, May 2000, pp. 268–279.

[5] J. F. Naughton and S. Seshadri, "On estimating the size of projections," in *Proc. of the Third international Conference on Database theory*, Paris, France, 1990, pp. 499–513.

[6] R. Rosenfeld, "Optimizing lexical and n-gram coverage via judicious use of linguistic data," in *Proc. of Eurospeech*, 1995, pp. 1763–1766.

[7] L. Lamel and G. Adda, "On designing pronunciation lexicons for large vocabulary continuous speech recognition," in *Proc. of International Conference on Spoken Language Processing*, 1996, pp. 6–9.

[8] S. Huang and B. Weir, "Estimating the total number of alleles using a sample coverage method," *Genetics*, vol. 159(3), pp. 1365–1373, November 2001.

[9] G. Youmans, "Measuring lexical style and competence: The type-token vocabulary curve," *Style*, vol. 24, pp. 584–599, 1990.

[10] N. Chipere, D. Malvern, and B. Richards, "Using a corpus of children's writing to test a solution to the sample size problem affecting type-token ratios," in *Corpora and language learners*, ser. Studies in corpus

linguistics, S. B. G. Aston and D. Stewart, Eds. Amsterdam: John Benjamins, 2004, vol. 17, pp. 139–147.

[11] I. Nation and D. Beglar, "A vocabulary size test," *The Language Teacher*, vol. 31(7), pp. 9–13, 2007.

[12] M. Agustin and M. Terrazas, "Examining the relationship between receptive vocabulary size and written skills of primary school learners," *Journal of the Spanish Association of Anglo-American Studies*, vol. 31.1, pp. 127–147, June 2009.

[13] J. Read, *Assessing Vocabulary*. Cambridge University Press, 2000.

[14] Y. Attali and J. Burstein, "Automated essay scoring with e-rater v.2.0," *Journal of Technology, Learning and Assessment*, vol. 4(3), 2006.

[15] L. Rudner, V. Garcia, and C. Welch, "An evaluation of intellimetric essay scoring system," *Journal of Technology, Learning and Assessment*, vol. 4(4), pp. 1–22, 2006.

[16] C. Leacock and M. Chodorow, "C-rater: Scoring of short-answer questions," *Computers and the Humanities*, vol. 37, pp. 389–405, 2003.

[17] J. Bernstein, "Phonepass testing: Structure and construct," Menlo Park, CA, Tech. Rep., 1999.

[18] Z. Wang and T. Schultz, "Non-native spontaneous speech recognition through polyphone decision tree specialization," in *Proc. of Eurospeech*, 2003, pp. 1449–1452.

[19] K. Zechner, D. Higgins, X. Xi, and D. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken english," *Speech Communication*, pp. 883–895, 2009.

[20] C. J. Fillmore, "Individual differences in language ability and language behavior," in *On Fluency*, C. J. Fillmore, D. Kempler, and W. Wang, Eds. New York, NY: Academic Press, 1979, vol. 14, pp. 85–101.

[21] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97(1-2), pp. 245–271, 1997.

[22] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, March 2003.

[23] D. Koller and M. Sahami, "Toward optimal feature selection," in *Proc. of the 13th International Conference on Machine Learning*, 1996, pp. 284–292.

[24] I. J. Good, "Turing's anticipation of empirical bayes in connection with the cryptanalysis of the naval enigma," *Journal of Statistics, Computation and Simulation*, vol. 66, pp. 101–111, 2000.

[25] D. Jurafsky and J. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2000.

[26] J. Bunge and M. Fitzpatrick, "Estimating the number of species: a review," *Journal of the American Statistical Association*, vol. 88(421), pp. 364–373, 1983.

[27] A. Gandolfi and C. C. A. Sastri, "Nonparametric estimations about species not observed in a random sample," *Milan Journal of Mathematics*, vol. 72, pp. 81–105, 2004.

[28] L. A. Goodman, "On the estimation of the number of classes in a population," *Annals of Mathematical Statistics*, vol. 20, pp. 572–579, 1949.

[29] A. Shlosser, "On estimating the size of the dictionary of a long text on the basis of a sample," *Engineering Cybernetics*, vol. 19, pp. 97–102, 1981.

[30] W. W. Esty, "Estimation of the number of classes in a population and the coverage of a sample," *Mathematical Scientist*, vol. 10, pp. 41–50, 1985.

[31] W. W. Esty, "Estimation of the size of a coinage: A survey and comparison of methods," *Numismatic Chronicle*, vol. 146, pp. 185–215, 1986.

[32] L. Holst, "On birthday, collector's and occupancy and other classical urn problems," *International Statistical Review*, vol. 54, pp. 15–27, 1986.

[33] I. J. Good, *Probability and the Weighting of Evidence*. Charles Griffin, London, 1950.

[34] R. Lewontin and T. Prout, "Estimation of the number of different classes in a population," *Biometrics*, vol. 12, pp. 211–223, 1956.

[35] W. Esty, "A normal limit law for a nonparametric estimator of the coverage of a random sample," *The Annals of Statistics*, vol. 11, pp. 905–912, 1983.

[36] V. M. Kalinin, "Functionals related to the poisson distribution and the statistical structure of a text," in *Proc. of the Steklov Institute of Mathematics*, vol. 79, 1965, pp. 6–19.

[37] R. H. Baayen, *Word Frequency Distributions*. New York: Kluwer Academic Publishers, 2001.

[38] M. Baroni and S. Evert, "Testing the extrapolation quality of word frequency models," in *Proc. of Corpus Linguistics*, ser. The Corpus Linguistics Conference Series, P. Danielsson and M. Wagenmakers, Eds., 2005, vol. 1.

[39] A. Chao, "Nonparametric estimation of the number of classes in a population," *Scandinavian Journal of Statistics, Theory and Applications*, vol. 11, pp. 265–270, 1984.

[40] A. Chao and S. Lee, "Estimating the number of classes via sample coverage," *Journal of the American Statistical Association*, vol. 87, pp. 210–217, 1992.

[41] K. P. Burnham and W. S. Overton, "Estimation of the size of a closed population when capture probabilities vary among animals," *Biometrika*, vol. 65, pp. 625–633, 1978.

[42] W. Preston, "The commonness, and rarity, of species," *Ecology*, vol. 29, pp. 254–283, 1948.

[43] G. Zipf, *Human behaviour and the principle of least effort.* Cambridge, MA, Addison-Wesley, 1949.

[44] S. Evert, "A simple lnre model for random character sequences," in *Proc. of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles*, Louvain-la-Neuve, Belgium, 2004, pp. 411–422.

[45] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Transactions on Information Theory*, vol. 50(7), pp. 1469–1481, 2004.

[46] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Always good turing: Asymptotically optimal probability estimation," *Science*, vol. 302(5644), pp. 427–431, Oct 2003.

[47] A. Orlitsky, N. Santhanam, K. Viswanathan, and J. Zhang, "Modeling profiles instead of values," in *Proc. of the 20th Annual Conference on Uncertainty in Artificial Intelligence*, 2004, pp. 426–435.

[48] A. Orlitsky and N. P. Santhanam, "Speaking of infinity," *IEEE Transactions on Information Theory*, vol. 50(10), pp. 2215–2230, Oct 2004.

[49] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Always good turing: Asymptotically optimal probability estimation," in *Proc. of the 44th Annual Symposium on Foundations of Computer Science*, Oct 2003.

[50] P. Laplace, *Philosophical essays on probabilities (Translated by A. Dale from the 5th (1825) edition).* NY: Springer Verlag, 1995.

[51] A. Hodges, *Alan Turing: The Enigma*. Walker and Co., 2000.

[52] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40(3/4), pp. 237–264, December 1953.

[53] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley, 1991.

[54] N. P. Santhanam, A. Orlitsky, and K. Viswanathan, "New tricks for old dogs: Large alphabet probability estimation," in *Proc. of the Information Theory Workshop*, September 2007, pp. 638–643.

[55] P. Lennon, "Investigating fluency in EFL: A quantitative approach," *Language Learning*, vol. 3, pp. 387–417, 1990.

[56] H. Riggenbach, "Towards an understanding of fluency: A microanalysis of non-native speaker conversations," *Discourse Processes*, vol. 14, pp. 423–441, 1991.

[57] J. Kinkade, "Predictors of esl fluency ratings by native and non-native raters," M.S. thesis, University of Pittsburgh, Pittsburgh, USA, 1995.

[58] G. Mizera, "Working memory and l2 oral fluency," Ph.D. dissertation, University of Pittsburgh, Pittsburgh, USA, 2006.

[59] H. Daller and H. Xue, "Lexical richness and the oral proficiency of chinese efl students: A comparison of different measures," in *Modelling and Assessing Vocabulary Knowledge*, ser. Cambridge Applied Linguistics Series, H. Daller, J. Milton, and J. Treffers-Daller, Eds. Cambridge University Press, 2007.

[60] D. Malvern and B. Richards, "Investigating accommodation in language proficiency interviews using a new measure of lexical diversity," *Language Testing*, vol. 19, pp. 85–104, 2002.

[61] B. Laufer and P. Nation, "Vocabulary size and use: Lexical richness in l2 written production," *Applied linguistics*, vol. 16(3), pp. 307–322, 1995.

[62] P. Meara and H. Bell, "P-lex: A simple and effective way of describing the lexical characteristics of short l2 tests," *Prospect*, vol. 16(3), pp. 5–19, Dec 2001.

[63] F. Tidball and J. Treffers-Daller, "Exploring measures of vocabulary richness in semi-spontaneous french speech: A quest for the holy grail?" in *Modeling and Assessing Vocabulary Knowledge*, ser. Cambridge Applied Linguistics Series, H. Daller, J. Milton, and J. Treffers-Daller, Eds. Cambridge University Press, 2007.

[64] J. Kormos and M. Dénes, "Exploring measures and perceptions of fluency in the speech of second language learners," *Systems*, vol. 32, pp. 145–164, 2004.

[65] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech," *Journal of the Acoustical Society of America*, vol. 111(6), pp. 2862–2873, 2002.

[66] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learner's fluency by means of automatic speech recognition technology," *Journal of the Acoustical Society of America*, vol. 107(2), pp. 989–999, 2000.

[67] S. Yoon, "Automated assessment of speech fluency for L2 English learners," Ph.D. dissertation, University of Illinois, Urbana-Champaign, USA, 2009.

[68] N. Ambady, F. J. Bernieri, and J. A. Richeson, "Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream," in *Advances in Experimental Social Psychology*, M. P. Zanna, Ed. San Diego, CA: Academic Press, 2000, pp. 201–272.

[69] M. Richardson, E. Dominowska, and R. Ragno, "Predicting clicks: Estimating the click-through rate for new ads," in *Proc. of the Sixteenth International World Wide Web Conference*, 2007, pp. 521–529.

[70] B. Möbius, "Rare events and closed domains: Two delicate concepts in speech synthesis," *International Journal of Speech Technology*, vol. 6(1), pp. 57–71, 2003.

[71] S. Chaudhury, R. Motwani, and V. Narasayya, "Random sampling for histrogram construction: How much is enough?" in *Proc. of the 1998 ACM SIGMOD international conference on Management of data*, 1998, pp. 436 – 447.

[72] H. E. Robbins, "Estimating the total probability of the unobserved outcomes of an experiment," *Annals of Mathematical Statistcs*, vol. 39, pp. 256–57, 1968.

[73] D. McAllester and R. E. Schapire, "On the convergence rate of good-turing estimators," in *Proc. 13th Annual Conference on Computational Learning Theory*, 2000, pp. 1–6.

[74] A. B. Wagner, P. Viswanath, and S. R. Kulkarni, "Strong consistency of the good-turing estimator," in *Proc. of the IEEE Symposium on Information Theory*, 2006, pp. 2526–2530.

[75] E. V. Khmaladze, "The statistical analysis of large number of rare events," Department of Mathematics and Statistics, CWI, Amsterdam, Tech. Rep. MS-R8804, 1987.

[76] E. V. Khmaladze and R. J. Chitashvili, "Statistical analysis of large number of rate events and related problems," *Probability theory and mathematical statistics (Russian)*, vol. 92, pp. 196–245, 1989.

[77] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Springer, 1998.

[78] J. D. Deuschel and D. W. Stroock, *Large Deviations*. Academic Press, 1989.

[79] P. Dupuis, C. Nuzman, and P. Whiting, "Large deviation asymptotics for occupancy programs," *The Annals of Probability*, vol. 32(3B), pp. 2765–2818, 2004.

[80] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2005.

[81] S. Yoon, L. Pierce, A. Huensch, E. Juul, S. Perkins, R. Sproat, and M. Hasegawa-Johnson, "Construction of a rated speech corpus of l2 learners' speech," *CALICO Journal*, 2009.

[82] N. de Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavioral Research Methods*, vol. 41(2), pp. 385–390, 2009.

[83] S. Evert and M. Baroni, *zipfR: Statistical models for word frequency distributions*, 2008, r package version 0.6-5. [Online]. Available: http://zipfR.R-Forge.R-project.org/

[84] S. Tauroza and D. Allison, "Speech rates in british english," *Applied Linguistics*, vol. 11, pp. 90–105, 1990.

[85] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2009, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[86] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. IEEE ICASSP '89*, 1989, pp. 532–535.

[87] R. Ejzenberg, "The role of task structure in oral fluency assessment," in *Proc. Annual Meeting of the American Association for Applied Linguistics*, 1995.

[88] B. F. Freed, "What makes us think that students who study abroad become fluent?" *Second Language Acquisition in a Study-Abroad Context*, vol. 14, pp. 123–148, 1995.

[89] R. Towell, R. Hawkins, and N. Bazergui, "The development of fluency in early learners of french," *Applied Lunguistics*, vol. 17, pp. 84–119, 1996.

[90] A. van Gelderen, "Prediction of global ratings of fluency and delivery in narrative discourse by linguistic measure-oral performances of students aged 11-12 years," *Language Testing*, vol. 11, pp. 291–319, 1994.

[91] N. Iwashita, A. Brown, T. McNamara, and S. O'Hagan, "Assessed levels of second language speaking proficiency: How distinct?" *Applied Linguistics*, vol. 29(1), pp. 24–49, 2008.

[92] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97(1-2), pp. 273–324, Dec. 1997.

[93] J. A. Nelder, *Generalized Linear Models.* New York: Chapman and Hall, 1989.

[94] S. Bhat and R. Sproat, "Knowing the unseen: Estimating vocabulary size over unseen samples," in *Proc. Proceedings of the 47th Annual Meeting of Computational Linguists*, Singapore, Aug. 2009, pp. 109–117.

[95] S. Bhat and K. Church, "Variable selection for ad prediction," in *Proc. of the 2nd International Workshop on Data Mining and Audience Intelligence for Advertising*, Las Vegas, USA, 2008, pp. 45–49.